Watchlight AI

SAMPLE DELIVERABLE · THREAT MODEL REPORT

AI Agent Security Threat Model

Patient intake and billing AI agent systems. Conducted under the 12 Principles for Agent Runtime Governance.

PREPARED FOR	Acme Health
SECTOR	Healthcare Network (Illustrative)
ENGAGEMENT	AI Agent Security Threat Modeling Workshop
SCOPE	Patient intake and billing AI agents
DELIVERED	February 2026

01 Executive Summary

Key findings and recommendations from the threat modeling engagement conducted with Acme Health's AI platform and security teams.

OVERALL RISK POSTURE

Acme Health has deployed 14 AI agents across patient intake, billing reconciliation, and provider communication workflows. The engagement identified **23 distinct threats** against these agents, including **4 critical**, **7 high**, and **12 medium** severity findings. The most significant risks concentrate around credential exposure, uncontrolled tool delegation, and insufficient policy enforcement at the agent-to-PHI interface.

Top Three Priorities

1. Credential broker for agents accessing EHR System

CRITICAL

Agents currently hold long-lived OAuth tokens with broad scope. Replace with purpose-bound capability tokens issued per task with 15-minute TTLs.

2. Intent-based authorization for billing write operations

HIGH

Billing agents can modify invoices without intent validation. Introduce runtime policy evaluation that validates declared intent against the action before execution.

3. Delegation chain verification across orchestrator agents

HIGH

Orchestrator agents delegate tasks to worker agents without scope narrowing. Authority widens at each hop, violating Principle 3.

Engagement Overview

DURATION 2 days on-site	AGENTS ASSESSED 14 (across 5 workflows)
THREATS IDENTIFIED 23 (4 critical, 7 high, 12 medium)	CONTROLS RECOMMENDED 18 (mapped to 9 principles)

02 Agent Inventory and Scope

Agents assessed during this engagement, including their purpose, ownership, and systems they currently access.

AGENT ID	PURPOSE	OWNER	SYSTEMS ACCESSED	RISK
intake-assist-01	Patient intake form completion and pre-screening	Patient Services	EHR System, SMS Gateway, Internal CRM	HIGH
billing-reconcile-02	Claims reconciliation and invoice adjustments	Revenue Cycle	Billing System, Clearinghouse, Payment Processor	CRITICAL
provider-comm-03	Provider notification and message routing	Clinical Ops	EHR System, Messaging, Email	MEDIUM
auth-helper-04	Prior authorization request preparation	Utilization Mgmt	EHR System, Payer Portals, PDF Tools	HIGH
orchestrator-main	Orchestrates sub-agents across workflows	AI Platform Team	All agents, Policy Store	CRITICAL
clinical-summary-05	Generates patient visit summaries	Clinical Ops	EHR System, Clinical Notes DB	MEDIUM
denial-response-06	Drafts appeals for claim denials	Revenue Cycle	Billing System, Payer Portals	HIGH
scheduling-assist-07	Appointment scheduling and reminders	Patient Services	EHR System, SMS Gateway, Calendar API	MEDIUM

Full inventory (14 agents) available in Appendix A. Agents were identified via platform logs, developer surveys, and network-level discovery. Three agents were found operating outside the central AI Platform Team's governance scope and were recommended for immediate registration or decommissioning.

03 Threat Findings

Identified threats organized by severity. Each finding includes the affected agents, attack scenario, impact, and recommended controls.

T-01: Long-lived OAuth tokens with full Epic scope

CRITICAL

Affected: billing-reconcile-02, intake-assist-01, auth-helper-04, denial-response-06. **Scenario:** A compromised agent or container inherits an OAuth token valid for 24 hours with broad Epic read/write scope. **Impact:** Unauthorized PHI access, patient record modification, HIPAA breach exposure. **Control:** Replace with purpose-bound capability tokens (15-minute TTL, scope-locked to specific records).

T-02: Orchestrator delegates without scope narrowing

CRITICAL

Affected: orchestrator-main and all sub-agents. **Scenario:** The orchestrator passes its full authority to worker agents, which can then perform any action the orchestrator is permitted. **Impact:** Privilege escalation through delegation. **Control:** Cryptographic delegation chains with mandatory scope narrowing at each hop (Principle 3).

T-03: Billing write operations without intent validation

HIGH

Affected: billing-reconcile-02. **Scenario:** The agent can modify any invoice on the billing system without declared intent validation. An agent that drifted from its intended task could issue refunds or adjustments outside policy. **Impact:** Financial loss, audit findings. **Control:** Runtime policy evaluation that validates declared intent against the requested action (Principle 2, Principle 7).

T-04: Prompt injection via unsanitized patient-submitted forms

HIGH

Affected: intake-assist-01. **Scenario:** Patient-entered form fields are passed directly to the LLM as context. An attacker could inject instructions that cause the agent to exfiltrate data or take unintended actions. **Impact:** Data leakage, unauthorized actions. **Control:** Content-level policy-as-code validation (Principle 13) plus action-level governance barriers.

T-05: No audit trail of agent decisions

HIGH

Affected: All agents. **Scenario:** Current logging captures API calls but not declared intent, evaluated policy, or the delegation chain that led to each action. Forensic reconstruction would require significant manual effort. **Impact:** Compliance exposure, inability to investigate incidents. **Control:** Behavioral observability with execution lineage (Principle 9).

Complete list of 23 threats (including 12 medium severity findings and 4 informational observations) available in Appendix B.

04 Recommended Controls and Roadmap

Recommended controls mapped to the identified threats, sequenced by priority and implementation effort.

Prioritized Remediation Roadmap

PHASE 1 · WEEKS 1–2

Immediate Actions

- Rotate all long-lived OAuth tokens
- Inventory unregistered agents
- Deploy temporary rate limits on billing writes
- Enable behavioral logging for critical agents

PHASE 2 · WEEKS 3–8

Credential Broker

- Deploy credential broker service
- Migrate billing and auth agents
- Implement purpose-bound tokens
- Retire long-lived tokens

PHASE 3 · WEEKS 9–16

Runtime Governance

- Deploy ARG control plane
- Author policies for intake and billing
- Integrate with Epic gateway
- Enable intent validation

PHASE 4 · WEEKS 17–24

Delegation & Lineage

- Cryptographic delegation chains
- Scope narrowing enforcement
- Execution lineage graph
- Compliance reporting

Controls Mapped to Principles

CONTROL	ADDRESSES THREATS	PRINCIPLE	EFFORT
C-01 Capability token broker	T-01, T-08, T-11	P11 Tool Governance	Medium
C-02 Delegation chain enforcement	T-02, T-09	P3 Authority	High
C-03 Intent-based policy engine	T-03, T-12, T-14	P7 Runtime Policy	Medium
C-04 Content guardrails for patient input	T-04, T-10	P13 Injection Defense	Low
C-05 Execution lineage graph	T-05, T-06, T-17	P9 Observability	Medium
C-06 Agent registry with trust state	T-07, T-13	P1 Identity	Low
C-07 Kill switches and circuit breakers	T-15, T-18	P10 Safe Failure	Low
C-08 Human-in-the-loop for write ops	T-03, T-16	P6 Human-in-Loop	Medium

Full control mapping (18 controls across 9 principles) available in Appendix C. Implementation guidance, including specific technology recommendations for each control, is provided in Appendix D.

05 Next Steps

Recommended next steps to begin addressing the findings in this report.

IMMEDIATE (THIS WEEK)

Review the four critical findings with your security leadership team. Begin inventorying the three unregistered agents and planning their governance boundary. No architectural changes required; this is a visibility and decision exercise.

SHORT-TERM (2–4 WEEKS)

Execute the Phase 1 immediate actions: rotate OAuth tokens, deploy rate limits on billing writes, and enable behavioral logging for the critical agents. These actions do not require new infrastructure and begin reducing exposure today.

MEDIUM-TERM (1–6 MONTHS)

Begin Phase 2 credential broker deployment followed by Phase 3 runtime governance integration. These phases introduce the Agent Runtime Governance layer and address the majority of the high-severity findings.

LONG-TERM (6+ MONTHS)

Complete Phase 4 delegation and lineage work. Establish an ongoing governance review cycle. Expand scope to the remaining agent workflows that were not included in this engagement.

Follow-Up Support

This report includes 30 days of follow-up support from Watchlight AI for clarification questions on findings and recommendations. For implementation assistance beyond the workshop scope, contact us about advisory retainers or the Founding Design Partner Program.

CONTACT

Aldo Pietropaolo, CEO · `aldo@watchlight.ai` · `watchlight.ai/workshops`