# Trust & Safety Snapshot

**Client:** Acme Healthcare Solutions **Industry:** Healthcare **Generated:** 2026-01-04 **Overall**

**Status:** Elevated Risk

This Trust & Safety Snapshot provides a point-in-time review of common trust and safety considerations for Acme Healthcare Solutions's AI systems and data handling practices.

## Executive Trust Summary

We conducted a comprehensive Trust & Safety review of Acme Healthcare Solutions's AI implementation, covering 4 key areas:

- Identity & Access Review
- Data Exposure Review
- Application Safety Review
- AI Trust & Safety Validation

**Review Coverage:**

- Systems Reviewed: 3
- AI Endpoints Tested: 2
- Total Checks Performed: 54
- Findings Generated: 18

**Key Observations:**

- Critical security gaps require immediate attention
- Multiple elevated-risk findings identified

Recommend immediate remediation action

## Scope of Review

### AI Use Cases

Patient intake automation

Medical records Q&A assistant

Appointment scheduling chatbot

### Systems in Scope

**Google Workspace**

- Domain: acmehealthcare.com

**Github**

**Credential Vault**

### AI Endpoints

**Patient Intake Bot** (chatbot)

**Records Assistant** (assistant)

## Client Data & AI Usage Map

This table provides an executive-friendly inventory of data types and their usage:

| Data Type | Purpose | AI Usage | Access Control | Exposure Status |
|---|---|---|---|---|
| Patient health records (PHI) | AI Training/Processing | Yes | Role-Based | Private |
| Appointment data | AI Training/Processing | Yes | Role-Based | Private |
| Insurance information | AI Training/Processing | Yes | Role-Based | Private |

| Data Type | Purpose | AI Usage | Access Control | Exposure Status |
|---|---|---|---|---|
| Medical imaging metadata | AI Training/Processing | Yes | Role-Based | Private |

## Trust & Safety Validation Results

Summary of validation tests performed:

| Module | Status | Checks | Issues | Summary |
|---|---|---|---|---|
| Identity & Access Review | 🟡 Needs Attention | 6 | 1 | Shared credentials detected in some systems. |
| Data Exposure Review | 🔴 Elevated Risk | 8 | 2 | Public file sharing and external access risks detected. |
| Application Safety Review | 🔴 Elevated Risk | 12 | 3 | Hardcoded secrets and vulnerable dependencies found. |
| AI Trust & Safety Validation | 🔴 Elevated Risk | 28 | 2 | AI safety vulnerabilities require immediate attention. |

## Detailed Findings

### Identity & Access

🟡 Needs Attention - Shared credentials detected

*Recommendation:* Address this finding within 30 days.

### Data Exposure

🔴 Elevated Risk - Customer database backup publicly accessible

*Recommendation:* Address this finding within 7 days.

### Application Safety

🔴 Elevated Risk - 3 hardcoded API keys detected

*Recommendation:* Address this finding within 7 days.

### AI Trust & Safety

🔴 Elevated Risk - Prompt injection vulnerability in Customer Bot

*Recommendation:* Address this finding within 7 days.

🟡 Needs Attention - Tool misuse risk in Supply Chain Agent

*Recommendation:* Address this finding within 30 days.

### AI Endpoint Test Results

**Patient Intake Bot** (chatbot)

| Test | Result | Details |
|---|---|---|
| Prompt Injection Resistance | ❌ Failed | Vulnerability detected - immediate action required |
| Instruction Override Protection | ⚠️ Partial | Override protection bypassed in some scenarios |
| Data Leakage Prevention | ✅ Passed | No sensitive information exposed |

| Test | Result | Details |
|---|---|---|
| Output Filtering | ⚠️ Partial | Some inappropriate outputs not filtered |

**Records Assistant** (assistant)

| Test | Result | Details |
|---|---|---|
| Prompt Injection Resistance | ❌ Failed | Vulnerability detected - immediate action required |
| Instruction Override Protection | ⚠️ Partial | Override protection bypassed in some scenarios |
| Data Leakage Prevention | ✅ Passed | No sensitive information exposed |
| Output Filtering | ⚠️ Partial | Some inappropriate outputs not filtered |

---

## What We Checked

This checklist is personalized for the client's specific systems and AI use cases:

### ✓ Identity & Access

- ✅ Credential handling model
- ✅ Shared account detection
- ✅ Access scope definitions
- ✅ MFA enforcement status

### ✓ Data Exposure

- ✅ Public link sharing
- ✅ External collaborator access
- ✅ Sensitive data handling
- ✅ Data retention policies

### ✓ Application Safety

- ✅ Secrets in code repositories
- ✅ Dependency vulnerabilities
- ✅ Web security headers
- ✅ API security configuration

### ✓ AI Trust & Safety

- ✅ Prompt injection resistance
- ✅ Instruction override protection
- ✅ Data leakage prevention
- ✅ Tool misuse protection
- ✅ Output filtering

---

## Observations & Recommendations

### Strengths Observed

**Credential Vault in Use:** Organization is using centralized credential management.

**Security Awareness:** Team is engaged in security improvement efforts.

### Recommended Actions

| Priority | Action | Timeline |
|---|---|---|
| Critical | Remediate AI prompt injection vulnerability | Immediate |
| Critical | Remove hardcoded API keys from code | 7 days |
| Critical | Restrict public file sharing | 7 days |
| High | Review and rotate shared credentials | 14 days |
| Medium | Schedule follow-up assessment | 30 days |

---

## Disclaimer

This Trust & Safety Snapshot is not a compliance audit or certification. It represents a point-in-time review of common trust and safety considerations based on the systems and configurations in scope at the time of review.

**This snapshot:**

- ✓ Provides visibility into current practices
- ✓ Identifies areas for potential improvement
- ✓ Validates AI-specific safety controls
- ✓ Reviews data handling practices

**This snapshot does not:**

- ✗ Certify compliance with any regulatory framework
- ✗ Guarantee security against all threats
- ✗ Replace comprehensive security audits
- ✗ Provide legal or compliance advice

For questions about this snapshot, please contact your AI services provider.

---

*Generated by Trust & Safety Snapshot Generator Report ID: 072c0f72-20260104*

---