SOLVEWORKS

Enterprise Security Blueprint for OpenClaw AI Deployment

A zero-trust framework for executive AI assistants

Prepared for: Revaly

Prepared by: SolveWorks | solveworks.io

Date: February 2026

Classification: Confidential

Executive Summary

Your concerns are valid — and they're exactly why managed deployment matters

Your team raised important security concerns about deploying OpenClaw-based AI assistants for Revaly's executive leadership. **We want to be clear: those concerns are well-founded.**

The articles you cited describe real vulnerabilities in the *open, consumer-facing* OpenClaw ecosystem:

⚠️ **Bitdefender Report:** 135,000+ exposed agents with API keys, credentials, and sensitive workflows accessible to anyone.

⚠️ **The Verge:** Skills marketplace vulnerabilities allowing malicious extensions to exfiltrate data from unsuspecting users.

⚠️ **Hacker News / ClawHub:** 341 malicious skills identified, including credential harvesters and prompt injection payloads.

These findings are accurate. They describe what happens when AI agents are deployed without enterprise controls — public endpoints, unvetted marketplace skills, default-open permissions, and no oversight layer.

A SolveWorks managed deployment is fundamentally different. We deploy inside *your* infrastructure — on servers your IT team provisions and controls. We don't use the public ecosystem. We don't install marketplace skills. We don't expose agents to the internet. Every integration is custom-built, source-reviewed, and locked to the minimum permission scope required.

🛡️ **The difference:** The vulnerabilities cited apply to the open platform — the equivalent of running unvetted apps from the internet. Our approach is the equivalent of a locked-down enterprise MDM deployment: no app store, no public access, full audit trail, human approval for every outbound action.

Threat Landscape & Our Response

For each concern you raised, here's the specific control we apply

⚠️ Risk	✅ Our Response
ClawHub / Marketplace Risk 341 malicious skills discovered. Public marketplace allows unvetted code to access agent context, credentials, and user data.	**Zero Public Skills** We never install marketplace skills. Every skill is custom-built by SolveWorks, fully source-reviewed, and scoped to Revaly's specific workflows. No third-party code runs in your environment.
Agent Exposure / Misconfiguration 135,000+ agents publicly accessible with leaked API keys and sensitive data. Default configurations leave agents open to the internet.	**Private Deployment** No public endpoints. Agents are deployed within *your* infrastructure — on servers Revaly provisions and controls. No agent is discoverable or accessible from outside your authorized network perimeter.
Credential Theft Malicious skills designed to harvest OAuth tokens, API keys, and session credentials from agent runtime environments.	**Credential Isolation** All secrets stored in encrypted vaults, never exposed to the agent runtime. OAuth tokens are scoped to minimum permissions with automatic rotation. The agent never sees raw credentials.
Blast Radius of Compromise If one agent is compromised, attackers can pivot to other systems, escalate privileges, and access sensitive data across the organization.	**Sandboxed Runtime** Each agent runs in an isolated container with no lateral movement capability. Least-privilege networking means even a compromised agent cannot reach other internal systems.
Executive Permission Scope AI assistants with broad permissions could send emails, modify documents, approve transactions, or share sensitive information without oversight.	**Human-in-the-Loop** Read-only by default. Every outbound action — send, share, modify, approve, pay — requires explicit human approval. Executives maintain full control over what the assistant can do.

Security Architecture

Eight core controls that make enterprise OpenClaw deployment safe

1

No Public Skills

Only custom, audited skills built by SolveWorks. Zero marketplace dependencies. Full source code review before deployment. Every skill is purpose-built for Revaly's workflows.

2

Sandboxed Runtime

Each agent runs in an isolated container on Revaly's own servers — with its own network namespace. No shared memory, no shared filesystem, no ability to reach adjacent workloads.

3

Least-Privilege Integrations

Read-only by default. Each integration scoped to the minimum data and actions required. Calendar read ≠ calendar write. Email read ≠ email send.

4

Human-in-the-Loop

Every outbound action requires explicit approval. Send email? Approve. Share document? Approve. Modify record? Approve. No autonomous outbound actions.

5

Credential Isolation

Secrets stored in encrypted vaults (HashiCorp Vault or equivalent). Automatic rotation on configurable schedules. Minimum-scope OAuth grants. Agent runtime never sees raw tokens.

6

Audit Logging

Every agent action logged with timestamp, user, action type, target resource, and approval status. Logs are immutable, queryable, and alertable. Full audit trail for compliance.

7

Network Segmentation

Agents can only reach explicitly allowlisted endpoints. No blanket internet access. DNS-level filtering. Egress rules enforced at the network layer by Revaly's IT team — not just application config.

8

Version Pinning

No self-updating agents. Every version change goes through change management with review, testing, and approval. Rollback capability on every deployment.

🔒 **Defense in depth:** These controls are layered — and Revaly controls the infrastructure kill switch. Even if one control fails, the others contain the impact. A compromised skill can't reach the network. A network breach can't access credentials. A credential leak can't perform actions without human approval.

Platform Security Evolution

OpenClaw is actively hardening its core — 40+ patches in a single release, with daily updates continuing

The security concerns your team identified were based on reports from early February 2026. Since then, OpenClaw has responded aggressively. **Version 2026.2.12 alone patched 40+ vulnerabilities**, with three more releases (2.13, 2.14, 2.15) following in rapid succession. This is a platform that takes security seriously and ships fixes fast.

Key Security Patches (2026.2.12 – 2026.2.15)

SSRF Protection

Gateway now enforces explicit deny policies with hostname allowlists on all URL parameters — preventing agents from being tricked into accessing internal network resources.

Prompt Injection Defense

Browser and web content treated as "untrusted by default." Tool result details stripped during transcript compaction to prevent replay attacks from malicious web content.

Mandatory Browser Auth

Browser control routes now require authentication even on loopback. Auth tokens auto-generated at startup if missing — closing a local privilege escalation vector.

Session Hijack Prevention

Webhook endpoints now reject payload sessionKey overrides by default. Prevents attackers from hijacking agent sessions through crafted webhook payloads.

Sandbox Hardening

Skill sync destinations confined to prevent filesystem escapes. Directory traversal via frontmatter-controlled names is now blocked. Separate browser container bind mounts added.

High-Risk Tool Blocking

Dangerous tools (session spawning, gateway control) blocked from HTTP API by default. Fail-closed permission model when tool identity is ambiguous.

Webhook Hardening

Constant-time secret comparison prevents timing attacks. Rate limiting on webhook endpoints. Authentication bypass via loopback proxy trust patched.

Signed Packages

macOS packages now SHA-256 signed. Malicious bundled hook ("soul-evil") removed from codebase. Remote config tampering via unauthenticated API patched.

📈 **Release velocity:** 4 security-focused releases in 5 days (2.12 → 2.15). External security researchers actively contributing patches. This is a platform in rapid security maturation — and a SolveWorks managed deployment ensures you're always on the latest hardened version.

Proposed Pilot Design

A 30-day controlled deployment to prove security and value

We propose starting exactly where you suggested: a small, controlled pilot with 3–4 members of your Executive Leadership Team. This gives your security and IT teams full visibility into how the deployment works before any broader rollout.

🎯 Scope: 3–4 ELT Members

Hand-selected executives who will benefit most from AI assistance. Each receives a dedicated, individually configured assistant with permissions tailored to their role.

🔒 Read-Only Integrations Only

- **Calendar** — Read access only. View upcoming meetings, attendees, agendas. Cannot create, modify, or delete events.
- **Email** — Read access only. Summarize inbox, flag priorities, draft responses for review. Cannot send, forward, or delete.
- **SharePoint** — Read access to specific, pre-approved folders only. Cannot upload, modify, or share documents.

✋ Human Approval for ALL Outbound Actions

During the pilot, every action that sends data outside the agent is gated by explicit human approval. No exceptions. The executive sees exactly what will be sent and confirms before it happens.

🏗️ Dedicated Sandbox Environment

Pilot agents run on Revaly's own infrastructure, provisioned by your IT team — completely isolated from production systems. SolveWorks deploys remotely via secure access granted by Revaly's IT. Dedicated logging, monitoring, and alerting. Revaly can revoke SolveWorks' access at any time.

📊 Weekly Security Reviews

Joint weekly meetings between SolveWorks and Revaly's security team. Review all agent activity logs, discuss any concerns, adjust permissions and controls in real-time.

✅ Success Criteria & Kill Switch

- **Security:** Zero unauthorized data access, zero credential exposure, zero unintended outbound actions
- **Value:** Measurable time savings reported by pilot participants (target: 5+ hrs/week)
- **Kill switch:** Infrastructure-level control. Revaly can cut access at the server level — shut down agents, revoke SolveWorks' remote access, or power off the environment entirely. No waiting, no dependencies
- **Decision point:** Day 30 — joint review to determine go/no-go for expansion

📈 Clear Expansion Path

Pilot (3–4 ELT) → Full ELT (9 members) → Department leads → Broader rollout. Each phase requires explicit sign-off from Revaly's security team. Permissions expand only when you're ready. The pace is entirely in your hands.

Ongoing Security Commitments

What we commit to for the lifetime of the engagement

🔍 Regular Penetration Testing

Annual third-party penetration testing of the deployment environment. Results shared with Revaly's security team. Remediation SLAs for any findings: critical (24h), high (72h), medium (2 weeks).

🚨 Incident Response SLA

Dedicated incident response process. Initial acknowledgment within 1 hour. Root cause analysis within 24 hours. Transparent communication throughout. Post-incident review and remediation plan for every event.

📋 Quarterly Security Reviews

Formal quarterly review of agent activity, permission scopes, integration health, and emerging threats. Joint session with Revaly's security and IT leadership to maintain alignment.

Ongoing Security Commitments (continued)

🏛️ SOC 2 Alignment Roadmap

Our deployment practices align with SOC 2 Type II trust service criteria. We provide documentation mapping our controls to SOC 2 requirements for your compliance team's review.

👤 Dedicated Security Contact

A named security point-of-contact at SolveWorks for Revaly. Direct line for security questions, concerns, or incident reporting. Not a support queue — a person who knows your deployment.

🔄 Continuous Version Management

Updates are deployed via secure remote access, with Revaly's IT approval required before any changes reach production. Every update is tested in staging first. You're never exposed to known vulnerabilities — and never surprised by untested changes.

🤝 **Bottom line:** We don't just deploy and walk away. SolveWorks acts as your ongoing security partner — monitoring the threat landscape, applying patches, and adapting controls as the platform evolves. Your team stays informed, in control, and protected.

SOLVEWORKS

Let's build this the right way.

Security isn't a checkbox — it's the foundation. We're ready to design a pilot that your security team is confident in.

solveworks.io | hello@solveworks.io