# SECURITY BRIEF: OPENCLAW

August 2026

OpenClaw is an open-source agentic AI framework that connects frontier models to computer control, browser actions, email, and file access through a local gateway, real capability, but almost entirely outside centralized IT's view. Atlas rates **OpenClaw risk level as R5 - CRITICAL** and does not recommend it for enterprise production. Usage should stay confined to isolated, lab-style "agentic workbench" environments, with explicit senior risk oversight and no regulated or production data.

| R1 – MINIMAL | R2 – LOW | R3 – MODERATE | R4 – HIGH | R5 – CRITICAL |
|---|---|---|---|---|
| Governed Hosted Chat | **Governed SaaS Chat** | Constrained SaaS Agent | Privileged SaaS Agent | Unmanaged Local Gateway |

The R5 rating is architectural, not behavioral: OpenClaw's gateway runs locally with the same host privileges as the person who installed it, bypassing the identity controls, DLP, and audit logging that govern everything else on the network. That's a materially different risk profile than a privileged SaaS agent like Claude CoWork, which we rate R4 - HIGH. CoWork still runs inside a vendor's tenant, with an identity and admin layer to inherit, however incomplete. OpenClaw inherits nothing, which is why it sits a full tier higher, and why lab-only containment is the only workable posture today.

## Data Protection and Exfiltration Risk

OpenClaw's defining feature is also its core exposure: the gateway runs locally with the same privileges as the person who installed it, acting on files, browser sessions, and system commands with nothing external checking its work. Two points matter most for capital markets:

### Screening That Keeps Failing

OpenClaw's own registry, ClawHub, screens skills with automated scanners that attackers keep bypassing, so a published skill can still carry hidden instructions the model executes as commands, and the same is true of untrusted emails, documents, and synced folders OpenClaw reads.

### No Clean Audit Trail

Because the gateway runs under the user's own account, nothing in the log separates agent actions from user actions, so a client asked to reconstruct an incident after the fact has no clean way to do it.

## Governance, Auditability, and Defensibility

For institutional clients, the question isn't whether OpenClaw is capable, it's whether anyone could prove what it did after the fact. Right now, OpenClaw has two clear gaps:

### No Central Logging or Identity Layer

Logs stay local, with no central aggregation, tamper resistance, or clean way to separate what OpenClaw did from what the person did. There's no native SSO, SCIM, or RBAC layer, so for Entra ID-managed firms, revoking access means uninstalling software and hunting down credentials, not a normal deprovisioning step.

### Consumer Messaging as an Exit Path

OpenClaw installs quickly and routes through consumer messaging apps already on the machine (WhatsApp, Telegram, Slack, Signal), opening a direct exit path for firm data that DLP never sees. Combined with plaintext-stored credentials, that's a channel compliance can't monitor and can't easily shut.

### Atlas Guidance: Proceed w/ Caution

If compelled to explore OpenClaw now, confine it to lab-only use: contain the blast radius while it's still small enough to learn from safely, and don't let the missing audit trail and identity layer become something you have to explain after the fact.