



OpenClaw Security: Put Approval Before Agent Actions

Description

CSA’s advisory says autonomous AI agents such as OpenClaw can face unpatched flaws, weak access control, sensitive-data exposure, malicious skills and memory poisoning. It recommends least privilege, trusted skills, dedicated short-lived credentials, persistent logs and human approval for high-stakes or irreversible actions.

This guide is for an individual or small organisation considering an OpenClaw-style autonomous AI agent. Its purpose is to limit access and irreversible actions before connecting the agent to browsers, code, files, APIs or accounts. The answer comes first because the costliest mistake is usually taking the next irreversible step before the controlling condition is known.

Follow the branch that fits

- The device contains sensitive personal or financial data:** Do not install the open-source agent there.
- The agent needs tools or APIs:** Issue narrow dedicated credentials instead of reusing human administrator access.
- An action sends money, code, data or messages:** Require a system-enforced human approval.
- Compromise is suspected:** Rebuild from a known-good baseline including memory and downstream services.

The table is a triage tool, not a substitute for the underlying authority. It separates the reader’s situation from the action, so a general rule is not applied to the wrong person, property, business, journey or account.

Inventory every capability

A prompt cannot compensate for unknown tool access. The controlling position was checked against [CSA OpenClaw advisory](#).

Do this: List files, browser sessions, APIs, shell and messaging. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Use least privilege

Administrator access expands the blast radius.

Do this: Create a dedicated operating account. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Isolate credentials

Long-lived human tokens are difficult to attribute and revoke.

Do this: Use scoped, short-lived secrets from a secure store. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Verify skills

Third-party skills can execute code or instructions.

Do this: Check provenance, permissions and updates. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Gate irreversible actions

Prompt-only cautions can be bypassed. The related operating detail was also checked against [IMDA responsible deployment case study](#).

Do this: Enforce approval outside the model. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Persist attributable logs

Default temporary logs may vanish during recovery.

Do this: Store actions in a protected persistent location. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Test failure paths

A stated control may not trigger under prompt injection or tool error.

Do this: Run negative tests before real data access. This turns the rule into a dated record that another person can review, instead of leaving the outcome to memory or an informal message.

Two original tools for this decision

a capability-to-permission matrix for files, browser, APIs, code, money and messaging

This LBRD analysis applies each branch above to the reader's actual role, timing and evidence. Write the facts in separate columns, mark unknowns, and do not convert an estimate into a confirmed eligibility result.

a negative-test plan covering malicious skills, memory poisoning, prompt injection and missed approval

Keep the source, date checked, decision owner, deadline and supporting document in the same record. The value of this tool is not the template itself; it is the visible connection between the official condition and the action taken.

Stress-test the plan

A household agent that can read email, control a browser and use a saved bank session is an all-purpose high-impact identity. Splitting those capabilities and forcing approval before any transfer or external message reduces the blast radius.

The example is an analysis, not a promise that an authority, operator, provider or professional will reach the same result. Change one material fact at a time and re-run the decision. If the route depends on a date, amount, pass type, legal form, age, location or approved drawing, verify that field at the point of action.

Before acting

1. Do not install the open-source agent there.
2. Issue narrow dedicated credentials instead of reusing human administrator access.
3. Require a system-enforced human approval.
4. Rebuild from a known-good baseline including memory and downstream services.
5. Store actions in a protected persistent location.
6. Run negative tests before real data access.

Save the two official pages with the date checked. If an online form, price, timetable, clinic network or approval condition changes, the current official service must take priority over this explainer.

Recheck the evidence before the final step

Read the official material in the order the decision occurs. First confirm who or what is covered. Next confirm the effective date, threshold, location or document requirement. Then record any exception and the evidence for using it. Finally, verify the live submission, booking, payment or approval channel. This sequence prevents a valid rule from being attached to the wrong case.

A second reviewer should be able to reproduce the conclusion from the saved facts. If they cannot, the file is not ready: identify the missing field, return to the primary source and label any remaining uncertainty plainly.

Do not combine separate controls into one yes-or-no answer. Eligibility, cost, timing, approval, suitability and service availability can each have a different source and owner. A favourable answer on one field does not cure a failed condition on another. Keep the branches separate until every consequential field is confirmed.

Also distinguish the date a rule was announced, the date it takes effect and the date the reader must act. Where a future change is involved, write both the present process and the transition point. That makes it clear which instruction applies today and what must be checked again later.

Limits and escalation

Agent safeguards are evolving and implementation-specific. High-risk environments need professional security review.

Where the facts are disputed or the consequence is material, pause and ask the controlling authority or an appropriately qualified professional. Keep the answer with the documents used to make the decision.

Related LBRD guides

For the next adjacent task, read [Employment Agency Key Appointment Holder: Pass the MOM Checks](#). You may also need [HDB Doorbell Cameras and CCTV: Check Before Installing](#).

Date Created

14/08/2026

Author

vanessakoh