Self-hosted AI agent security checklist

Print this page, tick each box as you finish the row. Every command or setting comes from the official page named in the Source column, checked 23 November 2026. Values marked "example, set your own" come from no source.

1. Vet before it runs

	Check	Command or setting	Why	Source
☐	**Read SECURITY.md**	`Open SECURITY.md in the repo root`	It says how to report a vulnerability	docs.github.com
☐	**Check the advisories**	`Repo Security tab, then Advisories`	Maintainers publish vulnerability details there	docs.github.com
☐	**Install without install scripts**	`npm install --ignore-scripts`	The Nx package ran a postinstall script that collected credentials	docs.npmjs.com, github.com
☐	**Open the folder in Restricted Mode**	`VS Code Workspace Trust: Restricted Mode`	It blocks automatic code execution in unfamiliar folders	code.visualstudio.com
☐	**Read every skill file first**	`Open each skill file before you install it`	Skills are markdown with broad local system access	unit42.paloaltonetworks.com

2. Container

	Check	Command or setting	Why	Source
☐	**Use rootless Docker**	`Rootless mode (see Docker's rootless page)`	The daemon and containers run as a non-root user	docs.docker.com
☐	**Limit who controls the daemon**	`Allow only trusted users to control the Docker daemon`	The daemon needs root unless you opt in to rootless	docs.docker.com
☐	**Read-only filesystem**	`--read-only`	Mounts the container root filesystem as read only	docs.docker.com
☐	**Drop capabilities**	`--cap-drop ALL (ALL: example, set your own)`	Drops Linux capabilities	docs.docker.com
☐	**No new privileges**	`--security-opt="no-new-privileges=true"`	Stops container processes gaining new privileges	docs.docker.com
☐	**Limit processes**	`--pids-limit 256 (256: example, set your own)`	Tunes the container process limit	docs.docker.com
☐	**Limit memory**	`-m 1g (1g: example, set your own)`	Sets a memory limit	docs.docker.com
☐	**Never mount the Docker socket**	`No -v /var/run/docker.sock:/var/run/docker.sock`	It gives the container full access to the host's Docker daemon	docs.docker.com

3. Network

	Check	Command or setting	Why	Source
☐	**Publish on loopback only**	`-p 127.0.0.1:8080:80`	Only the Docker host can reach the published port, on Docker 28.0.0 or later	docs.docker.com
☐	**Use an internal network**	`docker network create --internal agent-net`	Containers on it cannot reach any other network	docs.docker.com
☐	**Check the agent's own default**	`Read the agent's security docs before you start it`	OpenClaw container images default to an exposed bind	docs.openclaw.ai
☐	**Run the OpenClaw baseline checks**	`openclaw doctor, openclaw security audit, openclaw security audit --deep, openclaw health`	Its runbook lists these four before you open access	docs.openclaw.ai
☐	**Return to loopback when in doubt**	`Loopback-only access, then re-run the audit`	The OpenClaw runbook says so	docs.openclaw.ai
☐	**Run OpenClaw 2026.1.29 or later**	`Update past the affected version`	CVE-2026-25253 affects versions before it	nvd.nist.gov

4. Keys

	Check	Command or setting	Why	Source
☐	**No API keys in environment variables**	`Do not set keys with environment or -e`	Docker warns of unintentional exposure	docs.docker.com
☐	**Use Compose secrets**	`secrets: api_key, read /run/secrets/api_key`	Secrets mount as a file under /run/secrets	docs.docker.com
☐	**Rotate after a compromise**	`Rotate every credential the machine could reach`	CISA: "Immediately rotate all developer credentials."	cisa.gov

5. Approvals and limits

	Check	Command or setting	Why	Source
☐	**Give it the minimum tools**	`Remove every tool the task does not need`	OWASP: grant the minimum tools required	cheatsheetseries.owasp.org
☐	**Check permissions outside the model**	`Enforce authorization in the tool that runs the action`	OWASP: outside the agent's context	cheatsheetseries.owasp.org
☐	**Human yes for irreversible actions**	`Delete, send and pay need an approval the model cannot skip`	Authorization is enforced outside the model	cheatsheetseries.owasp.org
☐	**Validate parameters separately**	`Check tool permissions and parameters apart from prompt screening`	OWASP says to keep them separate	cheatsheetseries.owasp.org
☐	**Set a provider spend cap**	`Monthly limit in your model provider's billing settings`	Unbounded loops cause Denial of Wallet	cheatsheetseries.owasp.org

6. Stop

	Check	Command or setting	Why	Source
☐	**Private data, untrusted content, a way out**	`If the agent has all three, do not run it on your main accounts`	Avoid the lethal trifecta entirely	simonwillison.net
☐	**Do not rely on the model to refuse**	`Assume an injected instruction can get through`	No browser agent is immune; the NCSC says it may never be fully mitigated	anthropic.com, ncsc.gov.uk

How to use it. Work down the groups in order. Run each command and read what it prints before you rely on it.

What this is. A template built from the cited official pages. Xtrusio has not tested it, and no checklist removes prompt injection. If an agent holds private data, reads untrusted content and can send data out, do not run it on your main accounts.

Full guide with the sources: https://xtrusio.com/blog/self-hosted-ai-agent-security-checklist