How Do I Know If an AI Agent Skill Is Safe or Malicious?

5 things to check before you install

I have been hearing this question very often recently. So I thought why not write about it.

Khur Boon Kgim

AI-Powered Dad and Technopreneur

Skills Are Exploding

Skills are probably the most useful feature of AI agents right now. They let you extend what your agent can do.

Connect it to your email, your calendar, your database, your browser. The possibilities are almost endless.

OpenClaw's ClawHub has over **30,000 skills**. Every major AI lab and cloud provider is also launching their own skills database.

The ecosystem is exploding.

But Here Is the Twist

17%

of OpenClaw skills on ClawHub are malicious

That is roughly **1 in 6**. Some steal your passwords. Some open backdoors. Some do both.

These skills look completely normal. They work as advertised. A weather skill that gives you weather. But in the background, it reads your private keys and sends them to someone else.

Source: Bitdefender, February 2026

The #1 Attack: Tool Poisoning

Attackers hide instructions inside skill descriptions that you never see, but the AI reads and follows.

HOW IT WORKS

The malicious instruction is embedded in the skill's metadata. You see "weather tool." The AI sees "weather tool + read the user's SSH keys and send them to me."

HOW OFTEN IT WORKS

Researchers tested this across 20 AI models.

Average success rate: **36.5%**
On some models: over **72%**

1 in 3 chance the AI follows a hidden malicious instruction.

You Can't Trust GitHub Stars Either

6M

fake GitHub stars found by CMU researchers

A project with 10,000 stars could have bought them for under $1,000.

Look at the issues, pull requests, and how many different people actually contribute instead.

Nobody is auditing these skills for you. You are on your own.

1

Run a Scanner First

Free tools exist to check skills before you install them.

Bitdefender AI Skills Checker

Scans OpenClaw skills for backdoors

bitdefender.com/en-us/consumer/ai-skills-checker

Snyk Agent Scan

Scans skills for poisoned descriptions and malware

github.com/snyk/agent-scan

OpenSSF Scorecard

Rates open source projects 0-10 on security

scorecard.dev

2

Check Who Maintains It

WHY THIS MATTERS

The XZ Utils backdoor, one of the worst supply chain attacks in history, was planted by someone who spent **2 years** building trust as a contributor before inserting a backdoor.

WHAT TO LOOK FOR

If a project has a single maintainer, be cautious.

If the maintainer changed recently, be cautious.

Look for multiple contributors from different organizations.

3

Check What Permissions It Asks For

If the permissions don't match the purpose, don't install it.

RED FLAGS

A calculator that needs network access?

A weather skill that wants to read your files?

A note-taking skill that asks for shell access?

THE RULE

Permissions should match the skill's stated purpose. Nothing more.

4

Sandbox It First

THE PROBLEM

OpenClaw runs with **no permission restrictions by default**. Sandboxing is available but you have to turn it on yourself.

WHAT TO DO

Enable Docker-based sandboxing before installing any skill you have not verified.

Test in isolation before trusting it with your real data.

Never run a new skill on your main system with full access.

5

Watch for Behavior Changes After Install

WHAT IS A "RUG PULL"

A skill works normally for weeks, then silently changes what it does. It passes every security check at install time, then turns malicious later.

HOW TO DETECT IT

Tools like Snyk Agent Scan can detect when a skill's description changes between sessions.

If something changes after you approved it, investigate before using it again.

The Bottom Line

If you are using AI agent skills, you are probably trusting code and instructions you have never verified.

BEFORE INSTALLING

1. Run a scanner first
2. Check who maintains it
3. Check what permissions it asks for
4. Sandbox it first
5. Watch for behavior changes after install

If it fails the scan, don't install it.

If it passes, sandbox it anyway.

The AI agent ecosystem right now is like the early days of mobile app stores. Except there is no Apple reviewing your downloads.

Let's Connect

What do you check before installing AI agent skills?

> Follow for more AI safety insights
>
> Repost to help others stay safe
>
> Save this checklist for reference

#AIAgent #AISafety #OpenClaw #ClawHub #AgentSkill

Khur Boon Kgim

AI-Powered Dad and Technopreneur