Beings

Before you run an AI skill: five questions

A checklist for owners and finance leads connecting agents to Xero, banking, HMRC or payroll

Dave Johnstone · 25 September 2026 · 2 min read

A skill is a set of instructions, often with scripts, that an agent runs with whatever access you give it. The safety sits in the controls around the agent, not in the file. Anthropic's own advice is to use skills only from sources you built yourself or got from Anthropic, and to audit every file in one before you use it.

1. What can it reach?

List the folders, accounts and connected apps the agent can touch while the skill runs.

Good looks like: the agent has its own identity, only the permissions the task needs, and credentials with the shortest possible lifetime (NCSC).

2. What stops it, other than the instructions?

A line in the skill saying "never pay anything" is a request, not a control. The NCSC's advice: do not rely on prompting alone.

Good looks like: a technical limit, such as a read-only account, a sandbox, or network access restricted to an allowlist of approved destinations (NCSC).

3. Where does a human say yes?

Before anything is filed, paid or submitted.

Good looks like: an approval the agent cannot get past, asked for rarely enough that people read it. Watching what an agent does (human-on-the-loop) is not the same as approving it first (human-in-the-loop). Anthropic found users approved roughly 93% of Claude Code's permission prompts.

4. Who has read it, and who owns it?

Good looks like: someone other than its author has read every file, scripts included; the version is pinned, so an update cannot change what runs without review; one named person is responsible (Anthropic, NCSC).

5. How would you know, and how do you stop it?

Good looks like: logs the agent cannot edit, and a way to halt it immediately (NCSC).

Sources

- NCSC, "Managing the cyber risk of agentic AI", blog, 20 August 2026. ncsc.gov.uk/blogs/managing-the-cyber-risk-of-agentic-ai
- Anthropic, "Agent Skills" and "Skills for enterprise", platform.claude.com, accessed 25 September 2026.
- Anthropic, "How we contain Claude across products", 25 May 2026. anthropic.com/engineering/how-we-contain-claude

The article behind this checklist: beings.com/the-problem-of-ai-skills/