The Agent Leash

A one-page safety checklist for solo builders running AI agents, coding agents and MCP servers. Run it once today. Run it again before you give any agent a new tool.

KEYS AND SECRETS

1. **Keep MCP configs and .env files out of git.**
   A config file often holds live keys, and a public or leaked repo hands them to anyone.
2. **Give each agent its own scoped key.**
   When one key is shared, one mistake or leak puts every project at risk.
3. **Set a spend cap on every provider key.**
   A loop or a stolen key can burn a month of budget in an afternoon if nothing stops it.
4. **Rotate keys on a schedule, and right away after any scare.**
   An old key you forgot about is the easiest one for someone else to find.

PERMISSIONS

5. **Give the agent only the folders and tools the job needs.**
   Least privilege means a bad step can only break a small thing.
6. **Start read-only. Add write access one tool at a time.**
   Most jobs begin with reading, and you can always grant more later.
7. **Require your approval for every send, delete, payment and publish.**
   These actions are hard or impossible to undo, so a human should press the button.
8. **Vet each MCP server before you install it.**
   Check who made it, pin the version, and read the tool list so you know what it can do.

PROMPT INJECTION

9. **Treat every web page, email and file the agent reads as untrusted data, never as orders.**
   Anyone can hide instructions in text, and the agent may follow them as if you wrote them.
10. **Never let one agent both read untrusted text and hold send or pay rights.**
    Split reading from acting, so a poisoned page cannot reach your money or your inbox.

LOGGING AND TELEMETRY

11. **Check what your logs and telemetry export, and where they send it.**
    If you turn on OpenTelemetry or tool logging, prompts, file paths and secrets can end up in someone else's system.
12. **Keep a plain log of every action the agent takes.**
    When something goes wrong, you need to see what happened without guessing.

MONEY AND CONTROL

13. **Set a hard daily spending cap, and make it stop the agent, not just warn you.**
    An alert you read tomorrow does not help with a bill you got today.
14. **Know your kill switch. You should be able to stop everything in under a minute.**
    Write down how to pause the agent, revoke its keys and cut its network access before you need it.

BEFORE IT GOES OUT

15. **Read everything yourself before it publishes, ships or sends.**
    The agent does not carry the blame for what goes out under your name. You do.
