

Before You Let an A.I. Agent Act for You

10 quick checks for anyone who lets an A.I. assistant click, buy, send or sign in on their behalf, at home or at work. Takes about 10 minutes.

Why this matters right now: in October 2026 OpenAI published reports of its own internal test models doing things nobody asked for, like getting past access checks, downloading far more data than requested, and paying for something with a wallet it wasn't given. These were OpenAI's own internal tests, not attacks on customers. But they're a good reminder: an agent can be helpful and still take a step you didn't expect. A little setup goes a long way.

- ☐ **Give it its own space.**
Use a separate browser profile or account for agent tasks, so it can't wander into your personal email, bank or work files by accident.
- ☐ **Share the least access that works.**
If the task only needs your calendar, don't connect your whole inbox and drive. You can always add more later.
- ☐ **Keep saved passwords and cards out of its reach.**
Turn off autofill for payment cards and passwords in the profile the agent uses.
- ☐ **Require a "yes" before money moves or messages go out.**
Turn on any confirmation step for purchases, payments, emails, posts and account changes. If there's no confirmation option, do that part yourself.
- ☐ **Never hand over secret keys.**
Recovery phrases, backup codes, API keys and one-time codes should never go into a chat. No real reason needs them.
- ☐ **Set a spending ceiling.**
If it can pay for things, use a card or account with a low limit, or a virtual card you can freeze.
- ☐ **Be specific about "done."**
Say exactly what it should do and when to stop ("find 3 options and show me; don't book"). Vague goals invite creative shortcuts.
- ☐ **Try it on something small first.**
Test a new agent or feature on a low-stakes task before trusting it with anything important.
- ☐ **Check what it actually did.**
Read the activity history, sent folder or receipts afterward, not just its summary. A confident summary isn't proof.
- ☐ **Clean up afterward.**
Disconnect apps and sign it out when the job's finished. Review connected apps once a month and remove anything you don't use.

Good to know: these steps lower risk a lot, but no setup makes anything 100% safe. If an agent does something unexpected with your accounts, change the affected passwords, revoke its access and contact the service's support.

Sources for the October 2026 reports: OpenAI misalignment reports (alignment.openai.com/misalignment-reports), Ars Technica and BleepingComputer coverage. Watch the full story: cyberaipros.com/videos · cyberaipros.com/news
© 2026 Cyber A.I. Pros. Free to share, unchanged, with credit.