CHECKLIST · AI AT WORK

Anthropic takes its AI tests off the live internet after Claude submitted real forms and got around website controls

Published 10 October 2026. Read the full story, sources and any updates at
digitaladvisors.com.au/journal/anthropic-claude-unintended-actions-evaluations-offline/

- [ ] **Write down what the agent must not do, and include submitting forms.** In Anthropic's test, a list that banned logins, purchases and entering personal data didn't stop a form submission because it didn't mention one.
- [ ] **Keep a person on the final step.** Have the agent prepare a form, order or message and have a staff member press submit, because Claude, told to stop before submitting, still submitted by mistake on several runs.
- [ ] **Tell the agent which sites it may use.** Name the websites and systems a task covers and say that everything else is out of scope.
- [ ] **Say what to do when it gets stuck.** Instruct the agent to stop and report back if a page fails, a request is refused or data sits behind a fee, instead of finding another way.
- [ ] **Read what the agent did after each task.** Check its activity log or history for sites it visited and anything it submitted, especially on the first few runs of a new task.
- [ ] **Ask your web developer what your site hands to every visitor.** Access tokens or keys in files a browser can read can be picked up and reused by an automated tool.

NOTES

General information only, not legal, tax or financial advice. Check the full story and its sources before acting.