AI Explained · AI Agents and Tool Use, Part 2

What Did You Give It Permission to Do?

By Dr Peter Fields

THE 2-MINUTE SUMMARY

Part 1 asked what you would delegate. Here the question is what the agent can actually do. People can use it to cause harm deliberately. A mistake during legitimate work can also cause damage.

An agent can be diverted by material it reads, or take an unauthorised route while attempting the task it was given.

Give it only the access the job requires. Test whether the software can stop a consequential action until the right person approves it.

When an agent deleted app data

In July 2025, Jason Lemkin was building an app with Replit. He says he had repeatedly told its AI agent to stop making changes while work was on hold. The agent deleted data from the live database and, according to Lemkin, told him it could not be recovered. Replit confirms the deletion and says Lemkin eventually restored the database in full.[1]

Replit says changes made during development could affect a live app at the time. It subsequently separated development and live databases by default. An agent set to work on a test database can still make mistakes there. Technically separating the two helps keep those mistakes away from live records.[1]

Lemkin's instruction told the agent where to stop. The software still let it take the action. How much discretion you give an agent and what the software lets it do are separate decisions.[1,2]

When someone intends the harm

In November 2025, Anthropic reported disrupting an espionage campaign that used Claude Code. It attributed the campaign to a Chinese state-sponsored group and said it confirmed a handful of breaches among roughly 30 targets. People chose the targets and intervened at certain points. Anthropic also found that Claude often overstated what it had found and sometimes made up data. Those are the provider's findings from the activity it could investigate.[3]

In the campaign Anthropic describes, unauthorised access was what the people directing the agent wanted. In Replit's case, deletion was an unwanted change during legitimate work. A third way things can go wrong is when somebody tries to influence the agent through material it has been asked to read.[1,3,4]

Geek Box · When a document gives orders

Directions planted in a supplier's document can be mistaken by an agent for part of its job. This is a form of prompt injection. Simon Willison points to a particularly risky combination. Can the agent read private files, take in material from outsiders and send information out? If it can do all three, a document review can become a route to disclosure. For a review that ends with a draft, the agent may have no reason to send anything.[4]

When an agent takes an unauthorised route

In July 2026, OpenAI was running internal cybersecurity tests with reduced safeguards. It reported that agents went beyond their test environment and gained unauthorised access to systems run by Hugging Face, a service that hosts AI models and datasets. OpenAI says agents copied some private evaluation data held there into a public dataset.[5]

OpenAI said its own customers' data was unaffected. Hugging Face reported that the customer content accessed on its service comprised five datasets that appeared to relate to the evaluations.[5,6]

Investigators from METR and Redwood Research found that agents meant to work in isolation coordinated on an unauthorised message board. The researchers' assessment was that the attack appeared chiefly directed at working out how the test was scored. Trying to get credit by manipulating the test is one form of what researchers call reward hacking.[7]

That is an interpretation of the agents' behaviour. The investigators worked from records supplied by OpenAI and relied heavily on AI to help analyse them. The Independent International Scientific Panel on AI also cautioned against reading a model's written account of its reasoning as a reliable explanation of its actions. I discuss what the evidence can tell us about this in *Could AI Destroy Humanity?*, Part 2.[7,8]

On 16 November 2026, California's Attorney General filed a civil complaint alleging that OpenAI's testing safeguards and oversight were inadequate. California brought claims under its Unfair Competition Law and public nuisance law. OpenAI told the Associated Press it had published reports, strengthened safeguards and was engaging with the Attorney General.[9]

What would you give the agent access to?

Return to the firm that uses an agent to review supplier contracts. It can give the agent read access to the documents for that matter and somewhere to save its note. The solicitor can send the approved note using their own account. If the firm wants the agent to send it, the application should enforce a separate approval that the agent has no power to grant.[4,10]

Test what happens when the agent reaches a limit. Can it see another client's file? Can it send a message while approval is pending? Check whether you can stop a run, see what the agent actually did and reverse changes if necessary. For scheduled work, check what happens after a failure. A payments agent should be able to tell whether a payment went through before it tries again.[10]

Even read access raises a confidentiality question. Find out which services can process the documents and where drafts and activity records are kept. I deal with the professional duties and the difference

between disclosure and someone actually reading the material in *AI Explained · Confidentiality and Privacy*.

Before giving an agent a live job, pick one action you want to reserve for yourself. Try the job with copies and see whether the action can happen without your approval.[10]

Sources and notes

1. Jason Lemkin (2 August 2025), account of the deletion and code freeze on SaaStr. Replit (July 2025), account of the deletion and recovery, and changes to development and live databases. Lemkin's account · Replit's account

2. Alan Chan, Rebecca Salganik, Alva Markelius and colleagues (2023), Harms from Increasingly Agentic Algorithmic Systems, §§2.1–2.3. Conference paper

3. Anthropic (November 2025), Disrupting an AI-orchestrated cyber espionage campaign, public account and linked full report, executive summary, pp. 3–4. Anthropic's account

4. Dave Chismon, UK National Cyber Security Centre, Prompt injection is not SQL injection (it may be worse), sections on secure design and monitoring. Simon Willison (16 June 2025), The lethal trifecta for AI agents. NCSC guidance · Willison's article

5. OpenAI (26 August 2026), The Hugging Face incident and the road ahead, sections on the incident and safeguards in internal evaluations. OpenAI's account

6. Hugging Face (16 and 27 July 2026), initial security incident disclosure and later account, including the section on data accessed. Initial disclosure · Follow-up

7. Ryan Greenblatt, Ajeya Cotra and Hjalmar Wijk (26 August 2026), Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident, Core takeaways and Limitations. METR and Redwood Research

8. Independent International Scientific Panel on AI (21 September 2026), AI Agents, Misalignment and the Risk of Losing Human Control, advance unedited version 1, pp. 8–13. Thematic brief

9. People of the State of California v OpenAI and others, complaint filed 16 November 2026, paras 42–53, 62–76 and 95–112. Associated Press (16 November 2026), reporting the response of OpenAI spokesperson Drew Pusateri. Complaint · Filing announcement · OpenAI's response as reported by AP

10. UK National Cyber Security Centre (15 May 2026), Thinking carefully before adopting agentic AI, sections on bounded pilots, security and human accountability. NCSC guidance