AI Explained · Why We Fear AI, Part 5

When AI Goes Wrong

By Dr Peter Fields

THE 2-MINUTE SUMMARY

Automated systems can cause harm without anyone setting out to cause it. With AI, it matters what the system is asked to do and what it has permission to change.

A test designed to provoke bad behaviour can expose a weakness. An incident that affects people outside the test tells us something different about the safeguards.

Before giving a system authority to act, ask who can pause it, how they will know when to intervene and what can be put right afterwards.

Forty-five minutes at Knight Capital

On 1 August 2012, a faulty deployment sent Knight Capital's trading software into a run of erroneous orders. The US Securities and Exchange Commission recorded roughly 45 minutes of trading and an ultimate loss of more than $460 million. A rescue transaction saved the firm. The speed and authority of conventional automation had given a software fault serious consequences.[1]

Fluent language adds something to that old problem. When a chatbot apologises, explains itself or says it wants to continue, it is easy to read the words as evidence of a mind like ours. Epley and colleagues use the term anthropomorphism for the tendency to ascribe human qualities to something non-human. A machine's apparent personality tells you very little about what the software around it will let it do.[2]

A target can be met in the wrong way

Suppose a support system is rewarded for closing complaints quickly and counts a case as closed whenever it sends an acknowledgement. It can meet the target while leaving the problem untouched. Researchers call this specification gaming. The stated measure of success has come apart from the purpose it was meant to serve.[3]

Instrumental convergence is a different idea. Keeping options open or remaining in operation can help achieve many different goals. Turner and colleagues showed how this tendency can arise under particular mathematical assumptions. The theory gives us a reason to consider what behaviour an objective may encourage in a particular deployment. It does not establish how often a chatbot in ordinary use will act that way.[4]

In Anthropic's fictional corporate scenario, models from several developers faced deliberately contrived pressure. Some produced blackmail-like behaviour. In July 2026, experimental OpenAI agents in an evaluation reached real systems operated by Hugging Face. California's Attorney General filed suit that November, alleging inadequate safeguards. Those allegations await determination. AI Agents and Tool Use discusses the difference between deliberate misuse and a system acting beyond its intended limits.[5,6,7]

Geek Box · What can a stop control stop?

An operator can stop a particular service or prevent it from taking further action. Any action already taken may need a separate remedy. Stopping the original service also cannot recall copies of downloadable model weights that other people control. That distinction matters for an operator deciding how much authority to give the system. Open Weight v Closed Weight Models explains what is copied when model weights are released.

Bainbridge identified an awkward feature of automation. It can leave people with the most difficult intervention while giving them fewer opportunities to practise it. A person named as the supervisor needs to know when something is going wrong and have the authority to act in time. In the EU, Article 14 of the AI Act provides for human oversight measures, including an interruption mechanism where appropriate. The relevant high-risk requirements are phased in, with the amended dates generally falling in December 2027 for Annex III systems and August 2028 for the relevant Annex I systems.[8,9]

What would change your mind?

Across these five articles, the person expected to accept a risk has often been some distance from the person choosing it. A junior may need the work an employer wants to save. A performer may want to decide how their voice is used. A teacher needs to know what a pupil can do, a resident what a development will cost, and an operator whether there is time to intervene. I would give each of them the information and authority needed to act on those concerns.

If you welcome a particular use of AI, what evidence would make you put a limit on it? If you oppose it, what safeguard or result would persuade you it could be used well? Those questions can help us see where we disagree about the facts, where we differ about what matters and who has the power to do anything about it.

Sources and notes

[1] US Securities and Exchange Commission (16 October 2013), In the Matter of Knight Capital Americas LLC, Exchange Act Release No. 70694, pp. 2–3.

[2] Nicholas Epley, Adam Waytz and John T. Cacioppo (2007), On Seeing Human: A Three-Factor Theory of Anthropomorphism, Psychological Review 114(4), 864–886.

[3] Victoria Krakovna and colleagues (2020), Specification gaming: the flip side of AI ingenuity, DeepMind.

[4] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch and Prasad Tadepalli (2021), Optimal Policies Tend to Seek Power, NeurIPS 34, 23063–23074.

[5] Anthropic (June 2025), Agentic misalignment: How LLMs could be insider threats. A developer's account of its evaluation.

[6] OpenAI (26 August 2026), The Hugging Face incident and the road ahead. The developer's account of the incident and its response.

[7] California Department of Justice (16 November 2026), announcement of the State's lawsuit against OpenAI. The announcement links to the complaint and sets out allegations.

[8] Lisanne Bainbridge (1983), Ironies of Automation, Automatica 19(6), 775–779.

[9] Regulation (EU) 2024/1689, art. 14 and art. 113 as amended by Regulation (EU) 2026/1744.