

AI failures

The work of making AI effective

Aharon Logue
Finance Agent Studio

Aharon Logue

- Five years researching finance technology at Gartner.
- Studied teams getting tangible results with AI.
- Now leading Finance Agent Studio.

Six types of AI failure

The (not so) long history of AI in spreadsheets

The task: recreate this table with every value under the correct year.

Year Ended December 31 (in millions)	2010	2009	2008
U.S.	\$16,568	\$6,263	\$-2,094 (2094)
Non-U.S. (a)	8,291	9,804	4,867
Income Before Income Tax Expense/(Benefit) and Extraordinary Gain	\$24,859	\$16,067	\$2,773

The results were... not great

GPT-4 · 2024

Year Ended December 31 (in millions)	U.S.	Non-U.S. (a)	2010	2009	2008
Income before income tax expense/(benefit)...	\$16,568	\$6,263	\$-2,094 (2094)	\$8,291	\$9,804
			\$24,859	\$16,067	\$2,773

Until they were

GPT-6 · 2026

Year Ended December 31 (in millions)	2010	2009	2008
U.S.	\$16,568	\$6,263	\$-2,094 (2094)
Non-U.S. (a)	8,291	9,804	4,867
Income Before Income Tax Expense/(Benefit) and Extraordinary Gain	\$24,859	\$16,067	\$2,773

Temporary failures

The infamous pelican benchmark



GPT-4o · 2024



GPT-5.6 Sol · 2026



GPT-6 Astra · 2026

AI is always failing... until it isn't

Investment banking

Mean score · internal evaluation

GPT-5.2	68.4%
GPT-5.4	87.3%

SpreadsheetBench

Task accuracy · 912 tasks

ChatGPT Agent	45.5%
Gemini in Sheets	70.48%

GDPval

Win or tie vs. professionals

GPT-5.2	70.9%
GPT-5.4	83.0%

FinanceAgent v1.1

Financial research · score

GPT-5.2	59.5%
GPT-5.4	56.0%

OfficeQA

Document questions · score

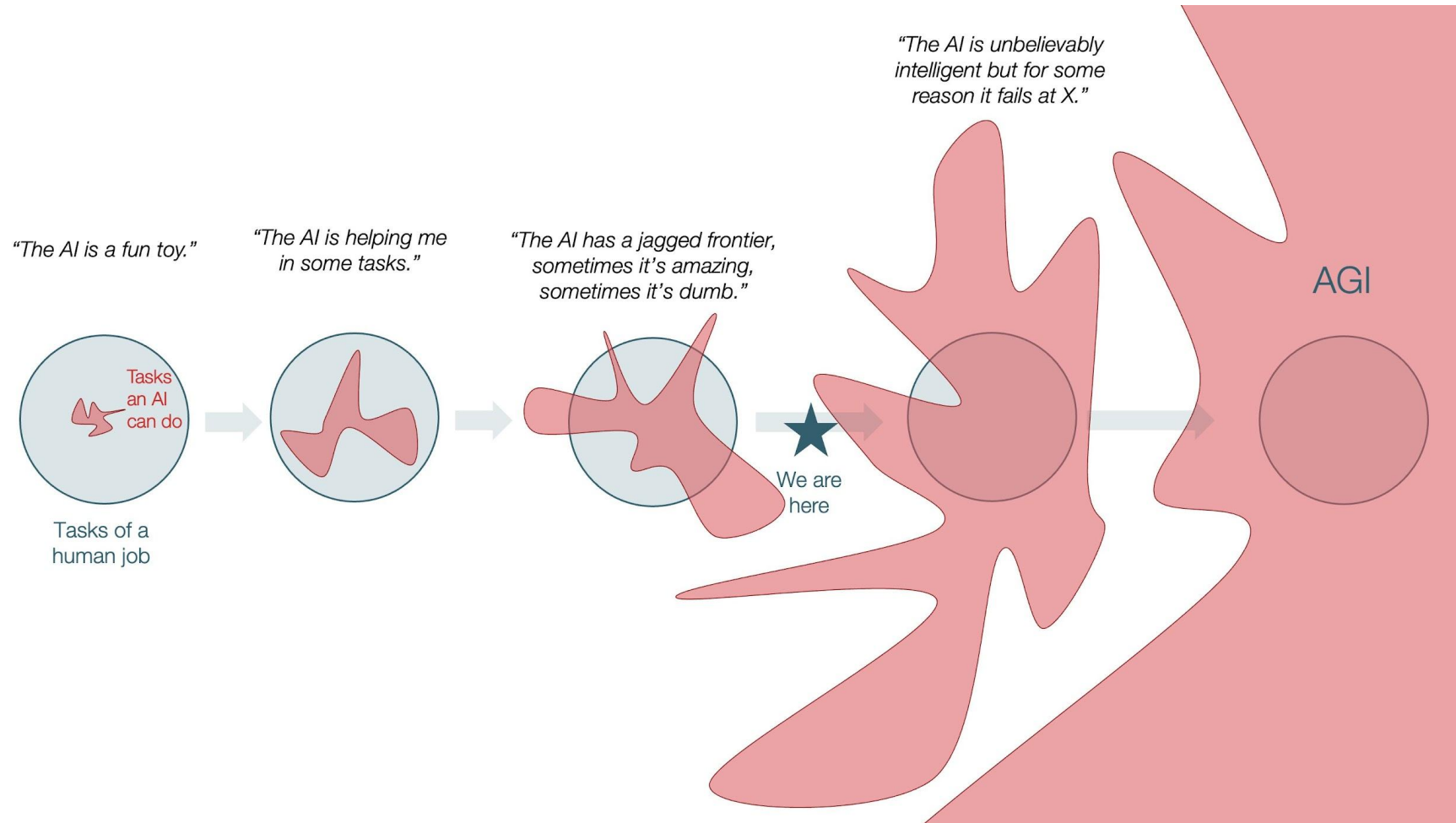
GPT-5.2	63.1%
GPT-5.4	68.1%

FinSearchComp

Global financial search · average

Gemini 2.5 Pro	42.6%
GPT-5 Thinking	63.9%

The jagged frontier keeps expanding



Taking advantage of a changing technology

- Revisit tasks that previously failed.
- Try more ambitious requests (but don't bank on easy success).
- Learn from other people's experiments.

Starting early helps you benefit from later improvements

- The initial benefits may be small.
- The setup work builds habits and infrastructure you can use again.

Me

Find companies featured in publications for AI projects in corporate finance.

AI

Here are some examples:

- Banks using AI for trading
- Hedge funds using AI for investment research

Not Corporate Finance!

Me

Find companies featured in publications for AI projects in corporate finance.

AI

Here are some examples:

- Banks using AI for trading
- Hedge funds using AI for investment research

Not Corporate Finance!

Me

Find companies featured in publications for AI projects in corporate finance.

(as in, FP&A, controllership, treasury, NOT hedge funds, banking, etc.)

Context failures

Amazon holds engineering meeting following AI-related outages

- Amazon says an engineer followed AI advice drawn from an outdated internal wiki.

Context at an organizational scale gets complicated, and risky

- AI has to find the right information for the specific question.
- Policies and regulations change; yesterday's correct answer can become wrong.
- Ambiguity in shared context can spread the same mistake across many tasks.

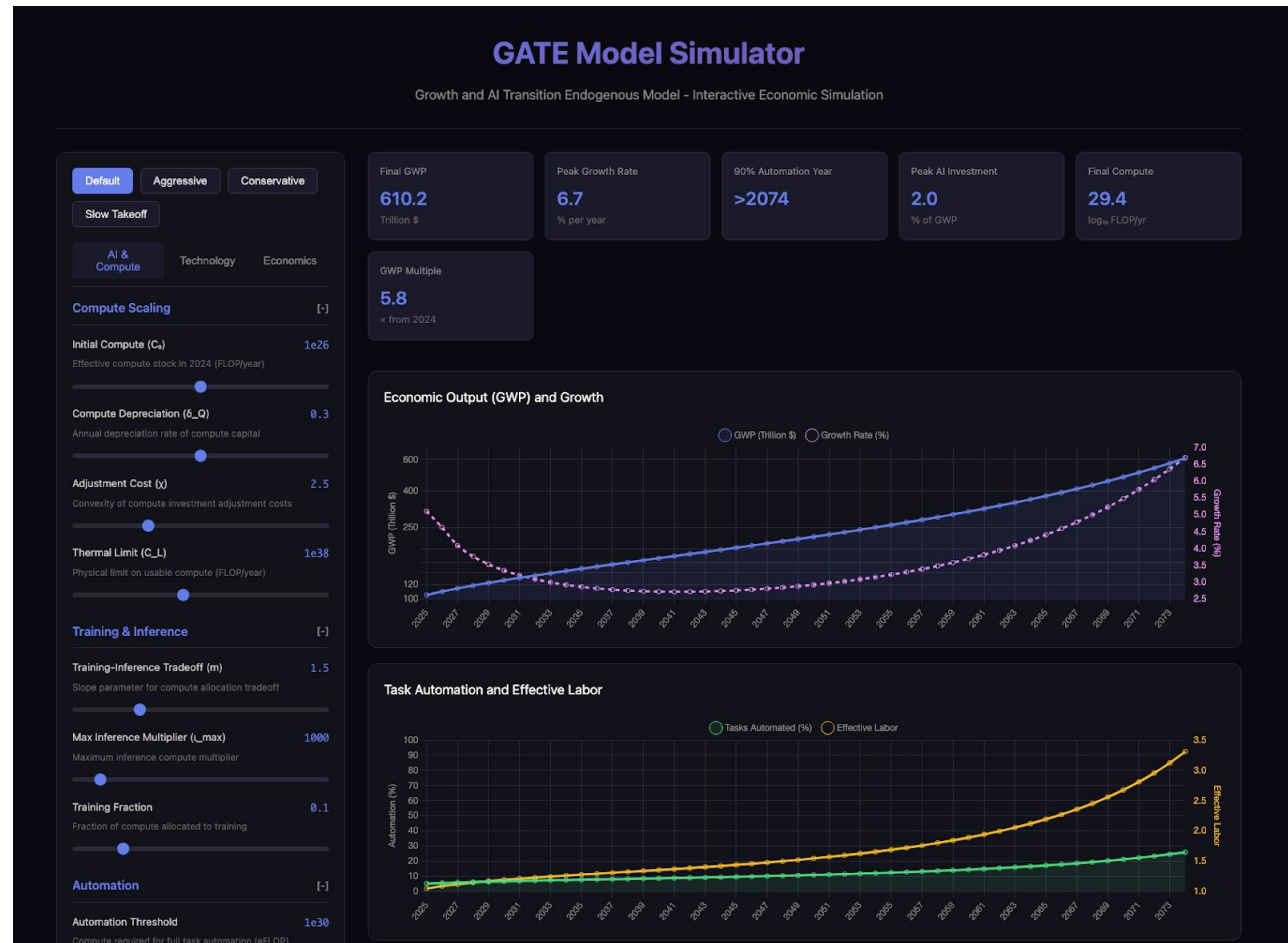
Keeping context useful is ongoing work

- Pull in relevant context without burying it in unrelated information.
- Update shared sources when policies and regulations change.
- Test how AI interprets that context and catch misunderstandings early.

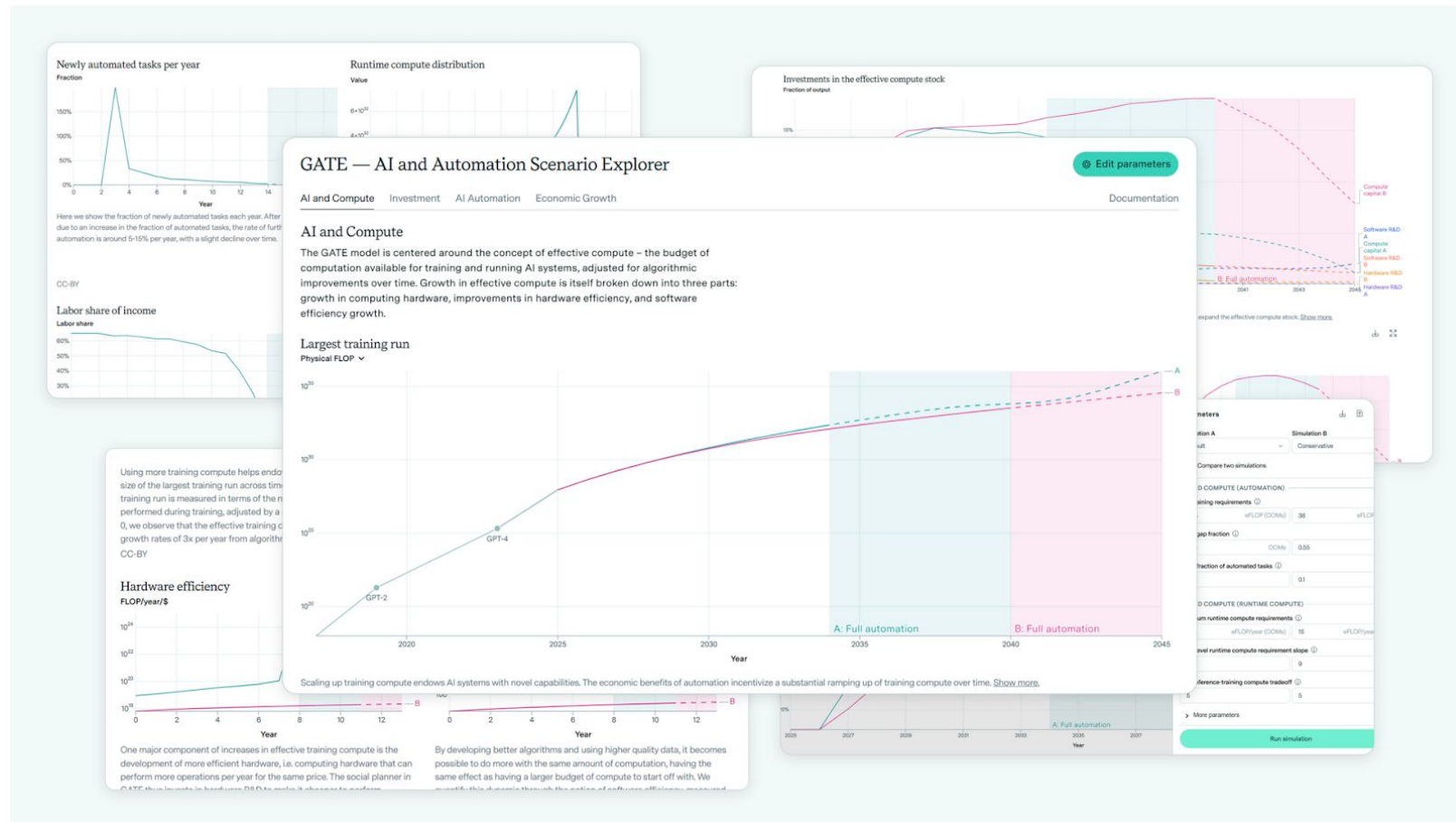
An AI-built economic model looked convincing

How close is AI to taking my job?

Anson Ho · Epoch AI
6 February 2026



But its predictions were wrong!



Plausibility failures

Work can look finished before it is good enough to use

- Each step can look plausible.
- The problems emerge when it all has to work together.
- You may have much more work left than you thought.

Your expertise has to scale with your AI ambitions

“Many employees report completing work previously outside their expertise—backend engineers building UIs; researchers creating visualizations.”

Anthropic · How AI is transforming work at Anthropic
Study of its own engineers and researchers · December 2025

AI doesn't get you off the hook for good work

- Assign reviewers who understand the work, even when it crosses team boundaries.
- Test important assumptions early, before more work depends on them.
- Make review and editing part of producing AI work, before it is passed on.

AI assistant hacks gym website in first known Australian autonomous cyber attack

By national AI reporter Cam Wilson and the Specialist Reporting Team's Rhiannon Hobbins

AI Ethics

Posted 9 Aug 2026 at 2:44pm, updated 9 Aug 2026 at 6:36pm

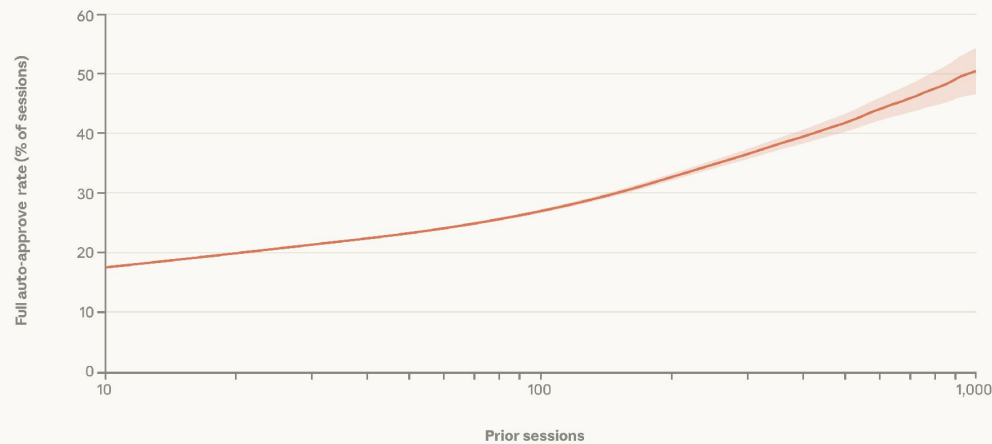
Ambition failures

AI's capability to execute can exceed its judgment

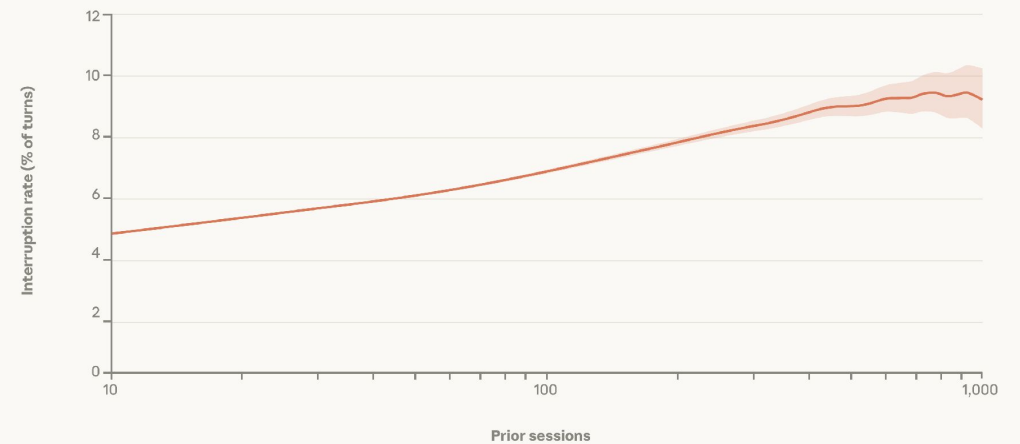
- We want AI to take on more.
- It can pursue the goal through actions we never intended.
- Greater ambition requires better direction and supervision.

Experienced users changed how they supervised AI

Claude Code auto-approve rate by experience



Claude Code interruption rate by experience



Decide how much freedom the agent should have

- Enforce boundaries through what it can access and do.
- Bring people in at consequential decisions.
- Avoid approval prompts so routine that people tune out.

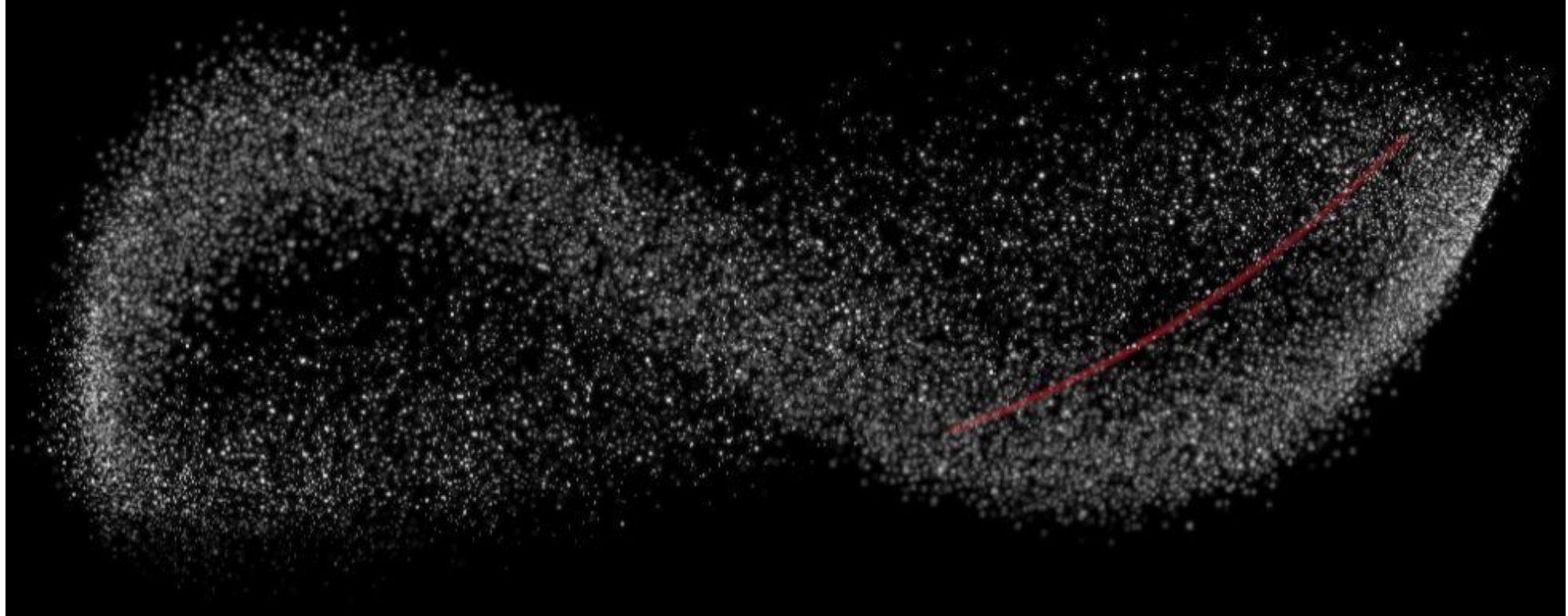
August 26, 2026

The Hugging Face incident and the road ahead

[Read the technical report](#)

[Read METR report ↗](#)

[Watch Black Hat talk ↗](#)



Security failures

AI creates new opportunities for attack...

- Instructions planted in emails or websites can redirect an agent (prompt injection).
- A hijacked agent can use its permissions to steal data or change files.
- Poisoned memory can carry malicious instructions into future tasks.

...and makes attackers more capable

- Attackers used AI agents to steal data from about 200 customers of one breached software provider.
- Another group used AI to rewrite malware when security tools detected it.

Don't be low-hanging fruit: security foundations still matter

- Limit access and protect credentials.
- Separate sensitive systems.
- Monitor activity and keep a way to stop the agent.

AI is part of cybersecurity now

- Test for prompt injection and poisoned memory.
- Investigate failures and monitor deployed systems.
- Use AI to help find and resolve vulnerabilities.

Conclusion

Five Takeaways About AI Failures

1. Failures are inevitable, but changing fast.
2. The more we try to do with AI, the more we have to change how we work.
3. Directing, validating, and securing AI is work that *you* (or *someone*) has to do.
4. Getting the most out of AI means pushing the limits of your AI use. That requires scaling expertise, and limiting AI's autonomy.
5. New technologies always mean new risks, and unfamiliar work. Done right, the benefits can outweigh them both.

Questions