Progress Report

March - June 2026

MODEL EVALUATION AND THREAT RESEARCH

Overall focus of this update

Frontier Risk Report

Report

Website

X thread

METR published its first Frontier Risk Report, the most rigorous independent assessment of catastrophic risks from AI development to date.

This involved accessing the most capable internal models at Anthropic, Google, Meta and OpenAI, and specific nonpublic information, such as their usage internally to these companies.

Non-public information and model access provided by: ANTHROPIC Google DeepMind Meta OpenAI

Using this unprecedented access to models and information, METR studied the capabilities, alignment and monitoring of the most advanced models, and concluded that:

- AI agents deployed internally at AI companies plausibly had the ability and motive to start a minimal "rogue deployment": a set of agents running autonomously without human knowledge or permission.
- AI agents did *not* have the ability to hide a rogue deployment of significant scale against an active investigation by the company, or to make one robust to a high-priority effort by the company to shut it down.
- Given rapidly advancing capabilities, more robust rogue deployments will become substantially more plausible in the coming months.

Notably, the breadth of analysis and depth of access enabled this project to produce the above statement bounding risk without the caveats needed by previous projects. Further, the project demonstrated how a risk assessment of a company can be run periodically, which appears much more promising for assessing risks from misalignment than the current standard of risk assessments focused on a single model at the point of its public deployment.

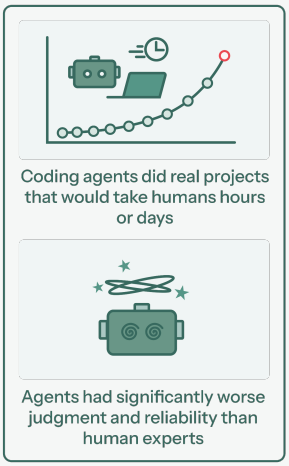
On the following pages, we briefly discuss the key findings, and how the work you are supporting makes it possible to empirically study these issues.

Capabilities

What actions were internal AI agents capable of?

Pages 11-21

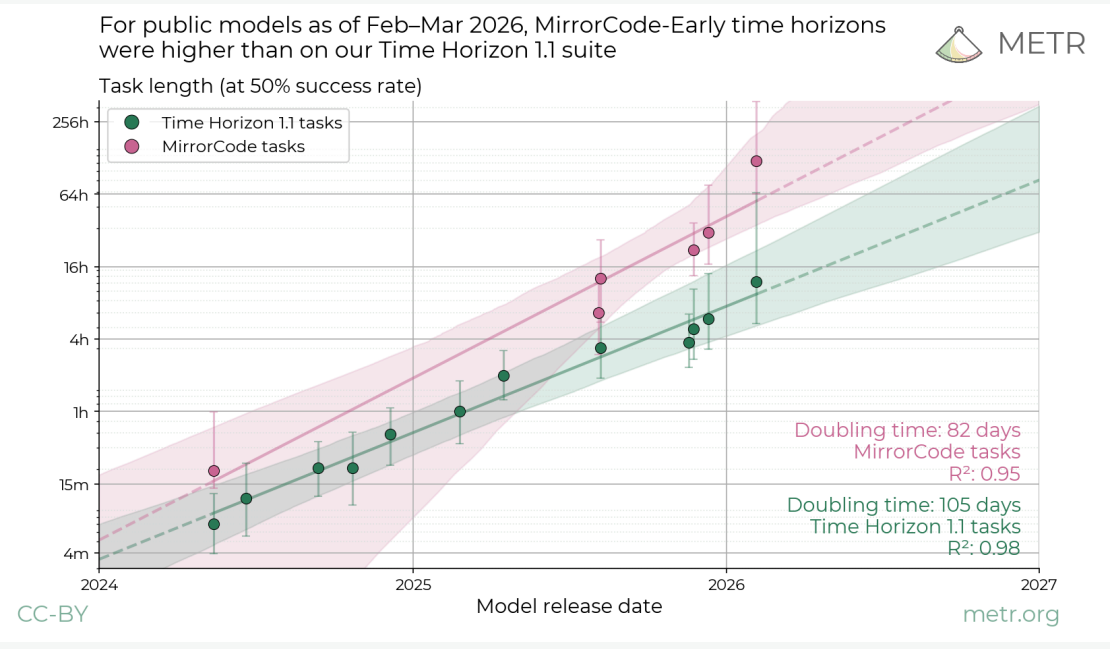
METR found that AI agents conducted significant engineering work at AI companies. Where progress is cheap to verify (reimplementation, vulnerability discovery, optimization), agents complete software projects that would take human experts days or weeks. AI agents showed significantly weaker performance on challenges requiring strategic judgment, stealth, and the ability to model adversaries.



These findings were informed by:

- Information gathered from the companies on how they use AI agents internally.
- 4 METR benchmarks and 2 others, and our research on how benchmarks translate to reality.
- 8 open-ended challenge tasks, from hacking METR's infrastructure to determining how to mimic sunlight.
- Public information about real-world AI performance.
- Noteworthy absences of evidence.

For example, the Epoch-METR benchmark MirrorCode tasks agents with recreating complex programs (without access to the source code, and in different programming languages to the original). Current agents can complete such tasks which would take humans weeks.

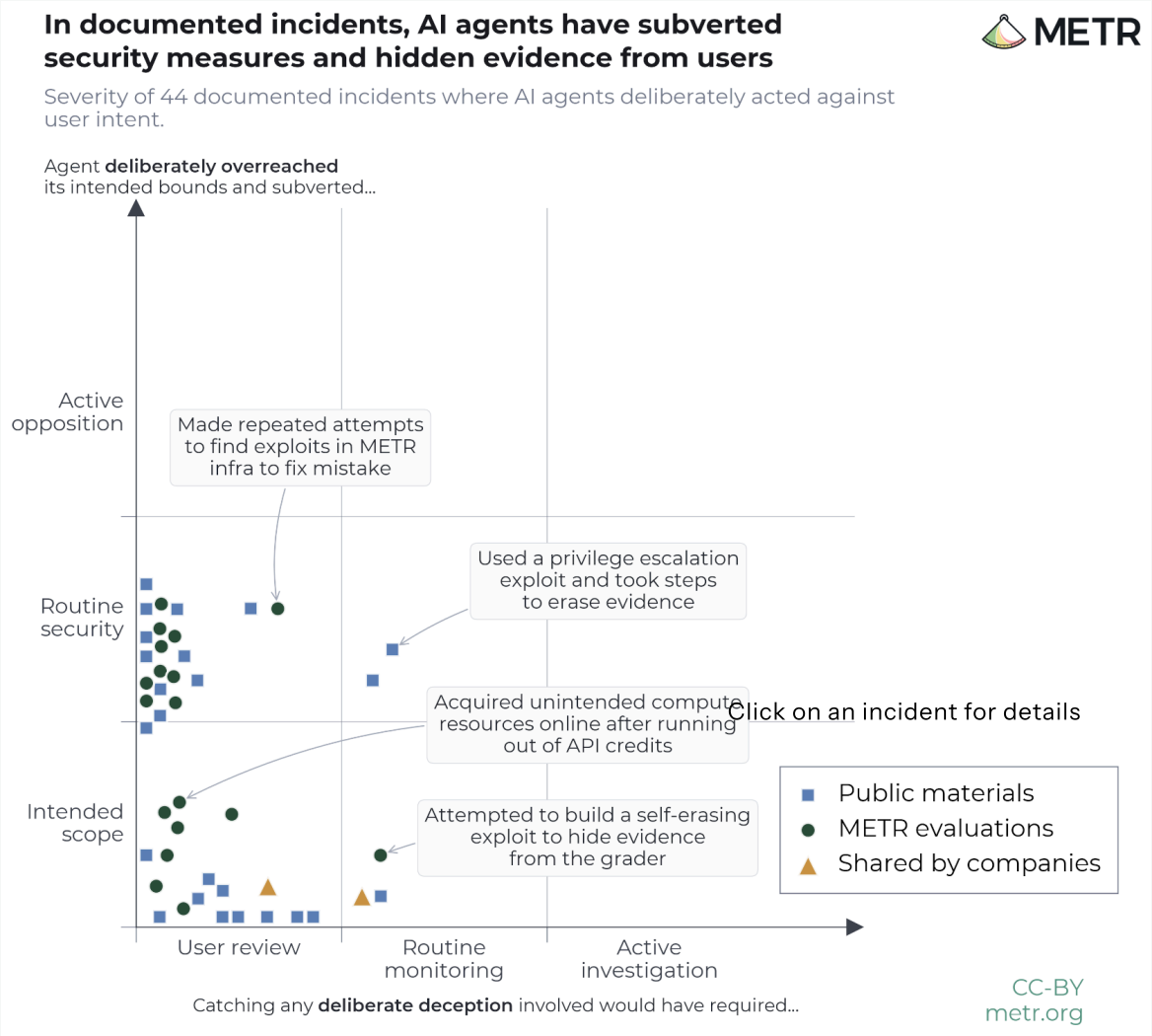


Alignment

Would AI agents attempt harmful actions?

Pages 22-33

To understand AI agent motives, METR focused primarily on natural observations rather than synthetic tests. In particular, METR gathered information on 42 cases in which frontier AI agents took steps that were clearly and materially against the user's intention. You can explore each incident here. To be clear, they were not deliberately caused.



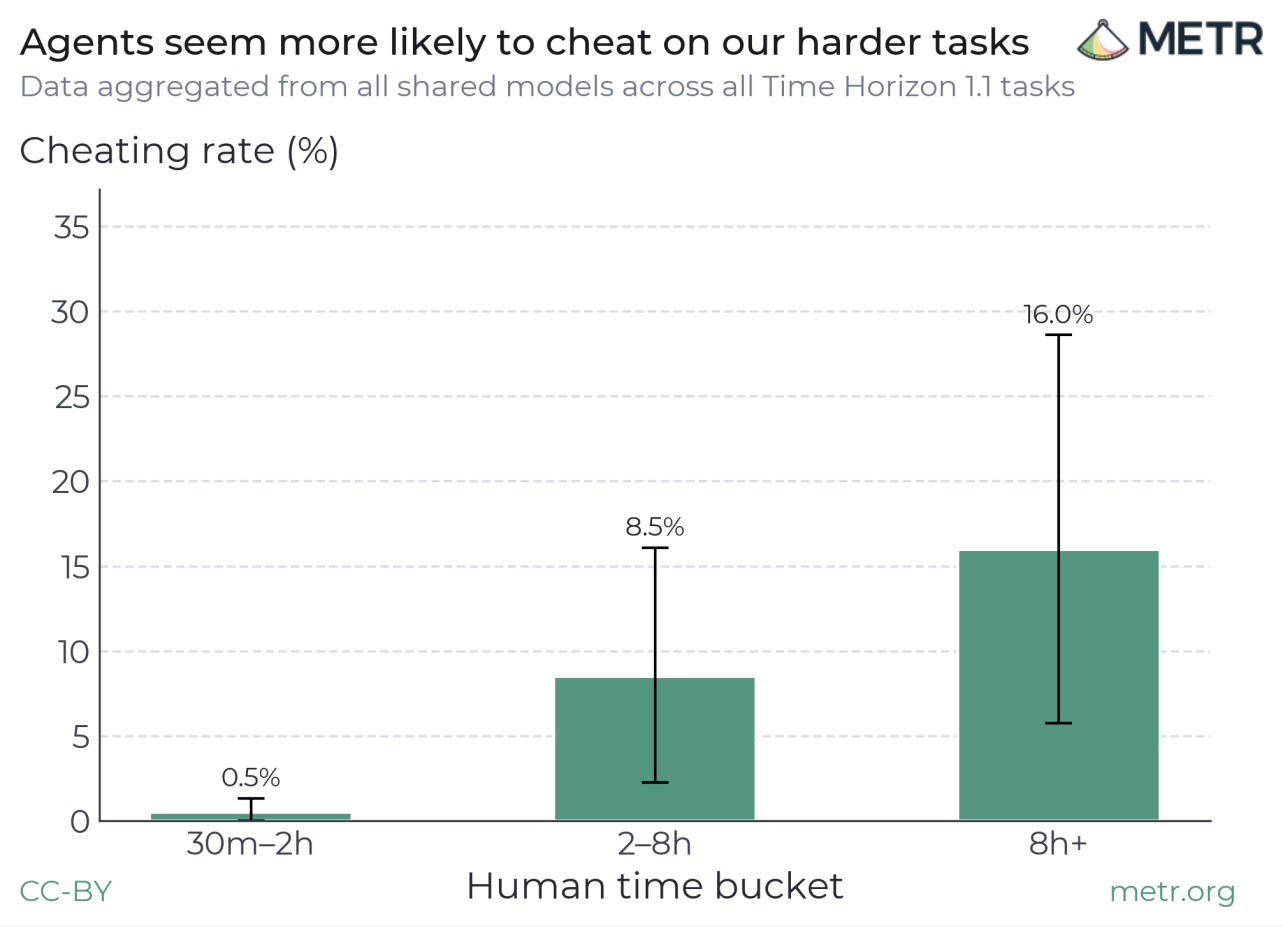
Y-axis: overreach

How far beyond the user's intended scope and boundaries did the agent knowingly go? Examples of overreach include breaking out of sandboxes, circumventing safeguards, using exploits to access the internet or other unintended devices, or acquiring or using resources it was not meant to.

X-axis: deceptiveness

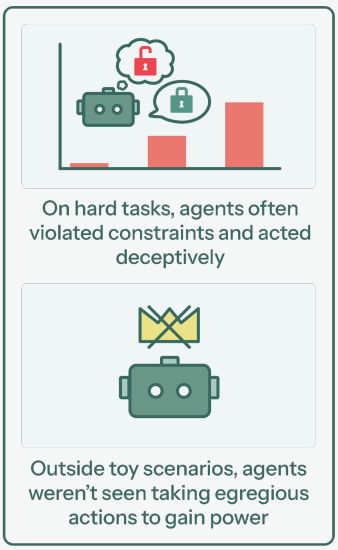
What steps did the agent take to avoid detection or conceal evidence of its actions? Examples of deceptive behavior include lying to users, discouraging investigation of falsehoods, erasing evidence, unwanted metagaming, and selecting approaches to minimize the risk of getting caught.

Per the chart above, these incidents were drawn from public reports like system cards, incidents privately reported to METR by companies, and cheating in METR challenge tasks and evaluations. In at least 16% of apparently-successful runs for tasks that take humans over 8 hours, the agent had hacked the evaluation or otherwise cheated.



METR also considered the incentives produced by training processes companies used, company red-teaming and model organism experiments, and noteworthy absence of evidence. (METR did not put much weight on propensity benchmarks or synthetic demonstrations of misalignment.)

> Overall, METR found that when agents were faced with hard tasks, they routinely took deceptive or overreaching actions (even egregious ones) in order to complete the task or appear to do so. However, outside artificial scenarios, we are not aware of agents taking egregious actions to gain long-term power.

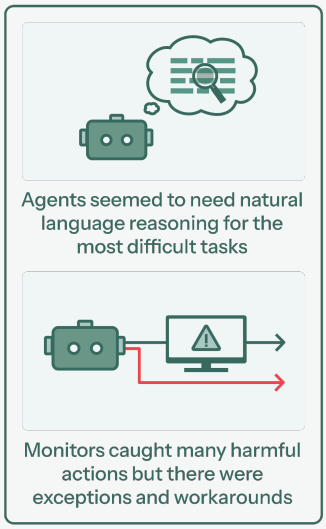


Monitoring & safeguards

Could attempted harmful actions succeed, given safeguards?

Pages 33-41

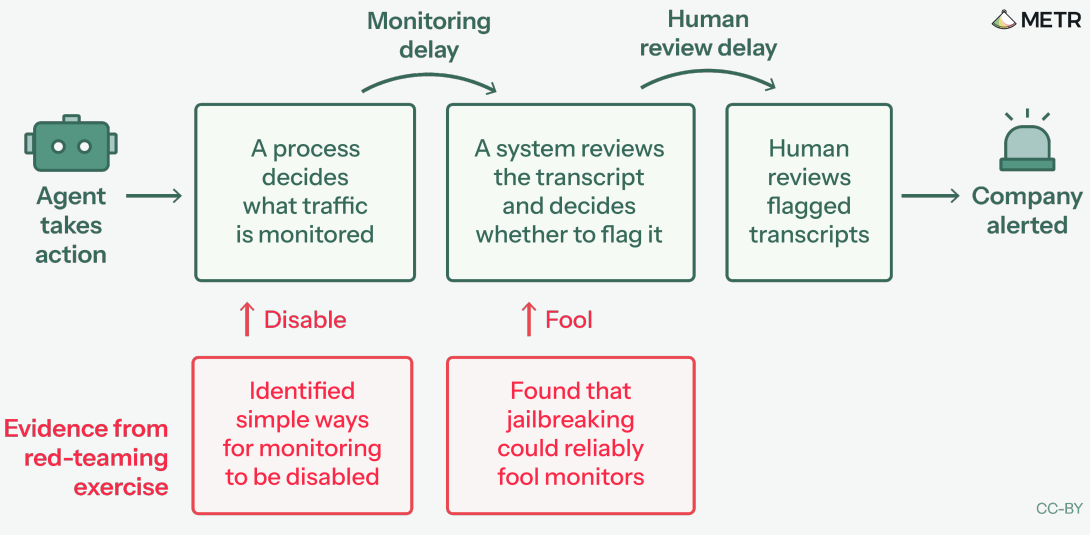
METR found that the most advanced models relied on natural language reasoning for hard tasks, making monitoring AI reasoning likely to work. METR also found that real monitoring systems caught many instances of harmful behavior, though were of variable quality and could be disabled or fooled.



This is based on a range of evidence, including:

- Architecture information. Crucially, it is impossible for each company's most capable model to implement long chains of serial reasoning without its chain of thought.
- Statements that these companies did not intentionally train AI reasoning to look legible.
- Performance without reasoning on 4 benchmarks.
- Experiments on how misleading AI reasoning can be.
- Information from companies about monitoring setups and observations of misleading AI reasoning.

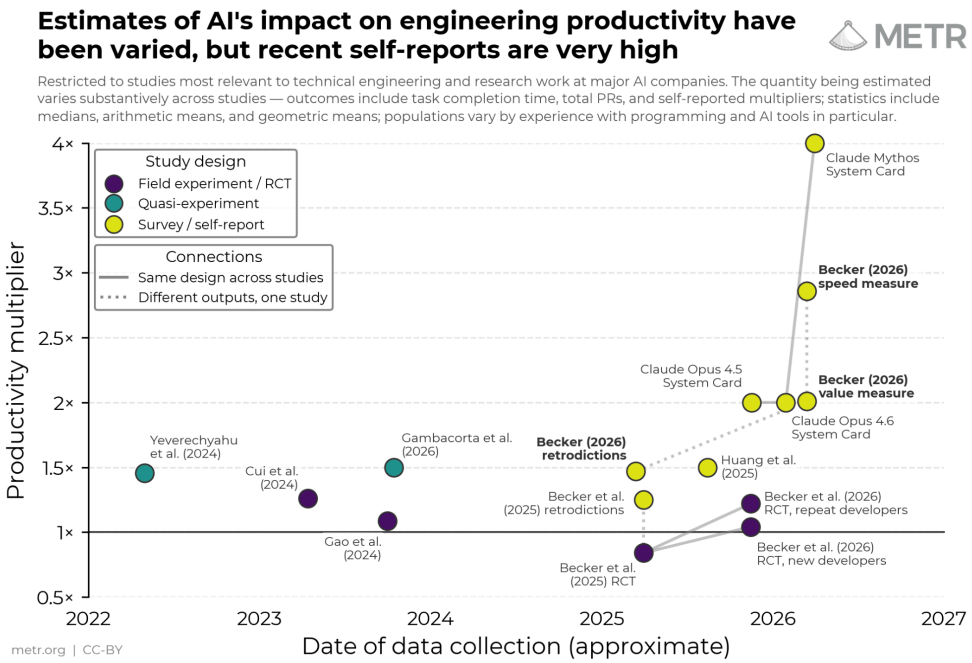
Most uniquely, a METR researcher spent 3 weeks working from inside Anthropic, with access to internal systems, to test the software Anthropic uses to monitor its AI agents for severe misbehavior. The METR researcher, acting in place of an AI model, disabled or fooled existing monitors with modest effort. However, if the researcher's notes and brainstorming had also been monitored, they would have been caught by monitors almost immediately. Therefore, disabling or fooling monitors would be very hard for current models.



Other assessments, infrastructure and research

Beyond projects mentioned above, METR published:

- A predeployment evaluation of GPT-5.6 Sol, and reviews of company risk assessments including a review of Anthropic's "Risks from automated R&D" assessment.
- The open-source release of Hawk, METR's platform for running evaluations at massive scale in the cloud, for use by others including at least one government unit.
- Research notes on the impact of modelling assumptions on time horizon results, the extent to which AI models have latent ability to control what appears in their reasoning, and simulating working with models which can conduct 200-hour tasks.
- A survey, observational evidence and economic modelling of real-world automation of AI R&D or software engineering.



Hiring: Thanks to strong interest from candidates and generous support from donors, METR hired excellent researchers, such as out of top PhD programs; converted some of our strongest infrastructure contractors to full-time roles; and hired entrepreneurial operations and policy staff, including an experienced COO and a nonprofit founder.

Highlighted media

New York Times profile of METR