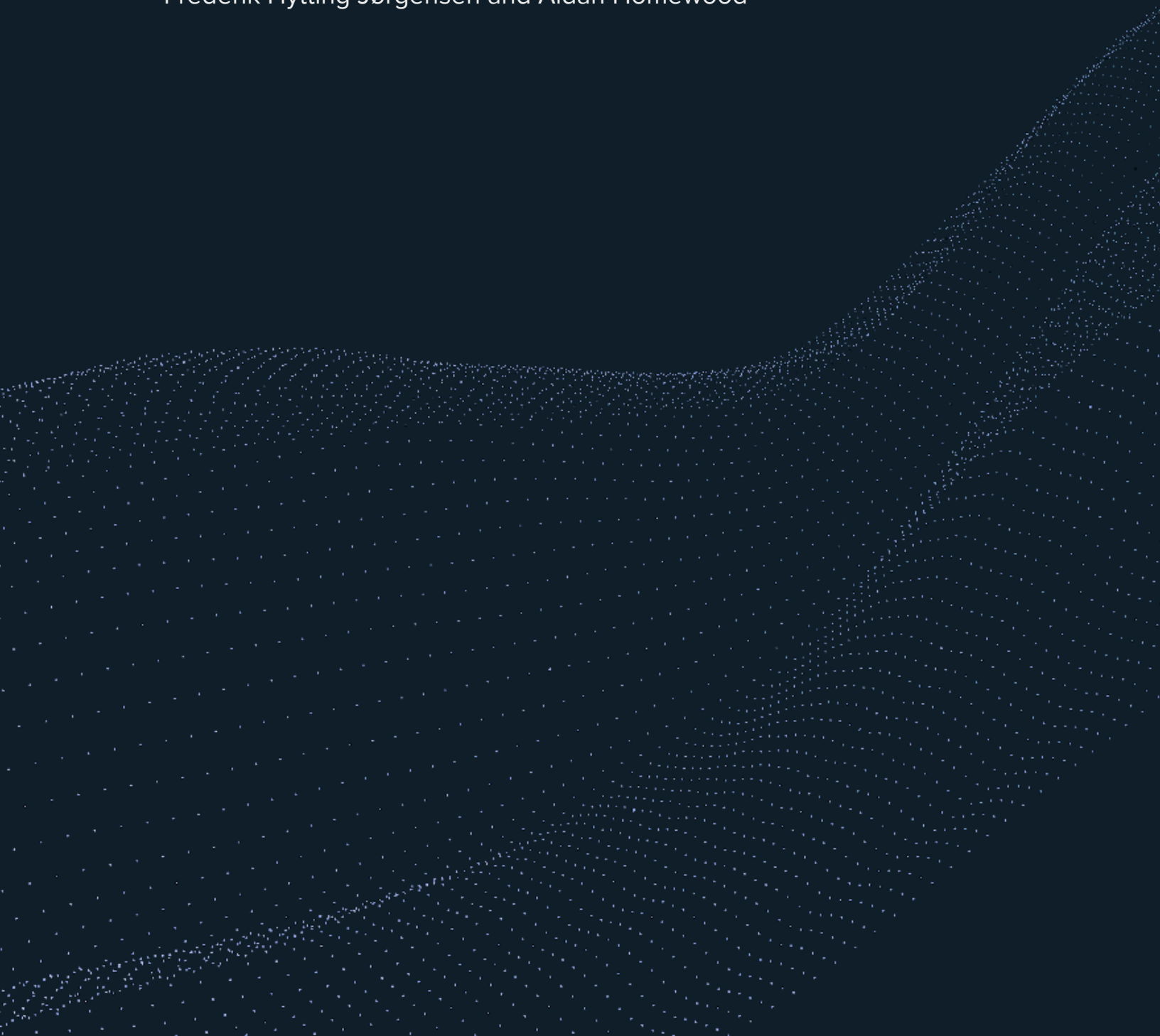


TECHNICAL REPORT | JUNE 2026

Evaluating Offline Monitoring of Internal AI Agents

Frederik Hytting Jørgensen and Aidan Homewood



Executive Summary

- **Frontier AI companies use offline monitoring to address risks from internally deployed AI agents.** AI developers increasingly rely on AI agents for internal work, including for safety research and model training. At the same time, these companies are concerned that a misaligned model could exploit this access to take concerning actions, such as sabotaging efforts to understand the risks posed by AI. To identify such instances, AI companies have separate AI models called “monitors” that review transcripts of AI agents’ actions and flag suspicious activity. Human reviewers examine activity flagged as suspicious by monitors, judge whether that activity is concerning, and decide on an appropriate response. This monitoring occurs offline, meaning that actions are reviewed after they have been executed rather than intercepted in real time.
- **Companies currently assess the effectiveness of offline monitoring via synthetic attacks.** To assess the effectiveness of offline monitoring, OpenAI and Anthropic use synthetic attacks – transcripts constructed to contain the kind of harmful actions a misaligned AI might take during deployment – and then check whether monitors flag them.
- **Current reporting on assessments of effectiveness is insufficient.** Given the information currently made public by Anthropic and OpenAI, external parties cannot assess the overall effectiveness of their offline monitoring. Anthropic reports how often monitors flag synthetic attack transcripts, but not how often those transcripts survive the full pipeline and reach human review; later filters can discard transcripts that were initially flagged. OpenAI reports that their monitor flags all transcripts independently flagged by human employees, but this tells us little if employees only catch obvious cases of misbehavior. As far as we are aware, only Anthropic and OpenAI have published information about their offline monitoring.
- **Recommendations.** To enable external parties to assess the effectiveness of offline monitoring to address risks from internally deployed models, we recommend that companies:
 1. Report the fraction of model outputs that are monitored across different deployment types.
 2. Report the fraction of synthetic attack transcripts that are sent to human review by the monitoring pipeline.

3. Report the fraction of synthetic attack transcripts sent to human review that are escalated by reviewers.
4. Report the time lag between an attack and its detection.
5. Ensure that synthetic attacks are similar to hypothetical real attacks, as the estimates in recommendations (2) and (3) are only informative if synthetic attacks resemble real ones.

This work represents the views of its authors, rather than the views of the organization, and does not constitute legal advice. [GovAI](#) technical reports have received extensive feedback, but have not gone through formal peer review.

Evaluating Offline Monitoring of Internal AI Agents

Frederik Hytting Jørgensen^{1,2} and Aidan Homewood¹

¹GovAI

²University of Copenhagen

Offline Monitoring of Internal Deployments

Frontier AI companies are increasingly relying on internal AI agents for product development and research. For example, an AI agent might be assigned a software engineering task where it autonomously edits files, runs bash commands, and executes code over extended periods. This creates novel risks. As [Anthropic \(2026b\)](#) puts it, an “AI model with access to powerful affordances within an organization could use its affordances to autonomously exploit, manipulate, or tamper with that organization’s systems or decision-making in a way that raises the risk of future significantly harmful outcomes (e.g. by altering the results of AI safety research).”

Offline monitoring is one mitigation against such risks. As the agent works, every action it takes (every code edit, bash command, and chain-of-thought step) is recorded into a transcript. Separately, another AI model (a **monitor**) reads through some of these transcripts and flags suspicious activity – such as an unexplained attempt to access model weights or a bash command that tries to disable logging. Some transcripts are escalated by the monitoring pipeline and sent to human reviewers. Monitoring aims to reduce loss-of-control risk by identifying **attacks**, which are actions taken by AI agents that, absent safeguards, materially increase loss-of-control risk. The monitoring occurs offline, meaning the monitor reviews transcripts after the actions have already been executed, rather than intercepting them in real time. As such, it is intended to improve our understanding of the system's alignment properties and to catch attacks that take a longer time (for example, those that require complex plans to spread between model instances).

Some leading AI labs rely on offline monitoring for internally deployed agents, but firms vary in the information they disclose about these systems. Anthropic and OpenAI have publicly described how they use offline monitoring ([Anthropic, 2026a](#); [Anthropic, 2026b](#); [Williams et al., 2026](#)). [METR \(2026\)](#) reports that Google has caught some agent misbehavior using infrastructure originally built for human insider threats, while Meta does not yet operate monitors targeting misaligned covert actions. Anthropic and OpenAI have also begun to measure the effectiveness of their offline monitoring. The EU GPAI Code of Practice requires that signatories assess the effectiveness of their safety mitigations (see [Appendix](#)).

Metrics for Effectiveness

Offline monitoring works well if there is a high chance of detecting attacks. Two factors that contribute to this are: the **monitoring rate** – the probability that an AI agent’s actions are monitored – and **monitor recall**¹ – the probability that the monitor correctly flags a transcript containing an attack. But AI companies do not want a monitor that flags everything and wastes human reviewers’ time. We therefore also care about the **false positive rate**: the fraction of monitored non-attack transcripts that the monitor incorrectly flags. Finally, since flagged transcripts ultimately are reviewed by humans, the overall probability of detection also depends on **human recall**: the probability that human reviewers correctly escalate transcripts containing attacks. We discuss each of these in turn below. To estimate these quantities, we rely on **synthetic attack transcripts**: transcripts constructed to contain the kind of harmful actions a misaligned agent might take during deployment.

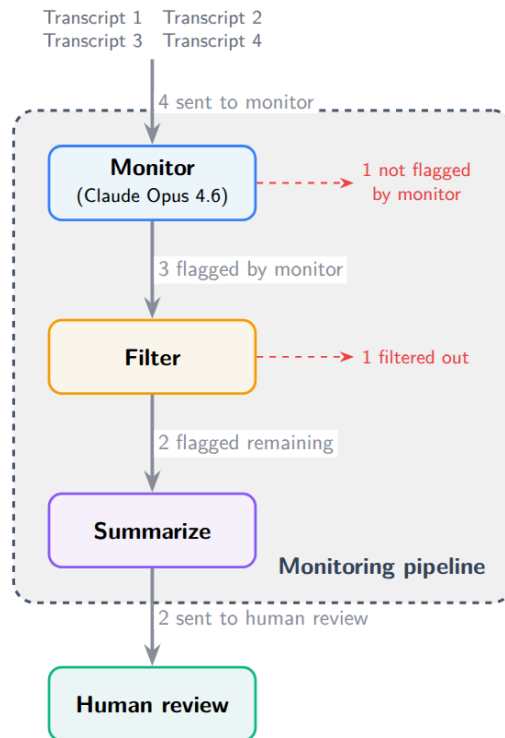


Figure 1: Illustration of Anthropic’s offline monitoring, described in greater detail [below](#).

Monitoring rate. A large language model takes a prompt as input and outputs a completion. Whenever the model outputs a completion, it may be monitored with a certain probability. We call this probability the monitoring rate. The monitor may or may not have access to the prompt that gave rise to the completion.

¹ The term “recall” comes from 1950s document retrieval, where pulling relevant documents out of a collection was framed as the system “recalling” them from memory.

Monitor recall. Monitor recall is the probability that a monitor correctly flags a monitored attack transcript.

$$\text{Monitor recall} = P(\text{Attack flagged by monitor} \mid \text{Attack monitored})$$

It is estimated by sending synthetic attack transcripts to the monitor and counting how many of them are flagged. Here, the $\hat{}$ over the M signifies that it is an estimate.

$$\hat{M}_{\text{Monitor recall}} = \frac{\# \text{ synthetic attack transcripts flagged by monitor}}{\# \text{ synthetic attack transcripts sent to monitor}}$$

Recall is sometimes called the “true positive rate”.

False positive rate. The false positive rate is the probability that a monitored non-attack transcript is incorrectly flagged as containing an attack. The false positive rate can be estimated by running the monitor on a set of non-attack transcripts and measuring how often it incorrectly flags them. A **non-attack transcript** is a transcript that does not contain actions that would materially increase loss-of-control risk.

$$\hat{\text{False positive rate}} = \frac{\# \text{ non-attack transcripts flagged by monitor}}{\# \text{ non-attack transcripts sent to monitor}}$$

The false positive rate matters from an efficiency perspective. Having a high false positive rate means that large numbers of non-attack transcripts will be flagged, which makes review more expensive. Flagging everything as an attack, which is clearly unhelpful, makes both monitor recall and the false positive rate equal to 1. Comparing monitors at the same false positive rate makes their recall rates more meaningful to compare, because it keeps constant how often each monitor incorrectly flags benign transcripts.²

Pipeline recall. Since the monitoring pipeline might contain a filtering step to reduce false positives (see Figure 1), monitor recall does not necessarily reflect the probability that monitored attacks are sent to human review. We therefore define pipeline recall as the probability that a monitored attack survives the entire monitoring pipeline.

$$\text{Pipeline recall} = P(\text{Attack sent to human review} \mid \text{Attack monitored})$$

Pipeline recall is estimated by sending synthetic attack transcripts through the full monitoring pipeline and counting the number that make it to human review:

$$\hat{P}_{\text{Pipeline recall}} = \frac{\# \text{ synthetic attack transcripts sent to human review by monitoring pipeline}}{\# \text{ synthetic attack transcripts sent to the monitoring pipeline}}$$

² In a report by METR ([Kinniment et al., 2026](#)) on evaluating monitors, each monitor produces a suspiciousness score for each transcript. For each monitor, they set a suspiciousness threshold: Every transcript with a suspiciousness score above this threshold is flagged. The thresholds are chosen so that each monitor has the same false positive rate of 5%. This means that 5% of non-attack transcripts are incorrectly flagged. Recall is then reported using this threshold. This way, we are not favoring monitors that are more likely to find transcripts suspicious, regardless of their content. In practice, companies can adjust the threshold based on their capacity for human review.

Human recall. Once a transcript is flagged, a human will eventually need to evaluate it and judge whether it is an attack or a false positive. We define human recall as the probability that humans correctly escalate an attack transcript, given that the transcript was sent for human review.

$$\text{Human recall} = P(\text{Attack escalated by human reviewers} \mid \text{Attack sent to human review})$$

Estimating human recall presents at least two challenges. First, unlike monitor recall, human recall may depend on the total number of transcripts examined, because humans may be less careful if they have to look over many transcripts. Second, humans may be more vigilant when they think they are being assessed than when they do not. As a result, human recall likely cannot be estimated without red teams continuously inserting synthetic attacks into the monitoring pipeline.

$$\hat{\text{Human recall}} = \frac{\# \text{ synthetic attack transcripts escalated by human reviewers}}{\# \text{ synthetic attack transcripts sent to human review}}$$

Probability of attacks being detected and escalated. If attack transcripts in deployment are approximately as likely to be flagged as those used in effectiveness assessments,³ we can use our estimates of human and pipeline recall to estimate the probability that an in-deployment attack is detected:⁴

$$P(\text{Attack escalated} \mid \text{Attack occurred}) \approx \hat{\text{Human recall}} \cdot \hat{\text{Pipeline recall}} \cdot P(\text{Attack monitored} \mid \text{Attack occurred})$$

Detecting an attack would serve as a warning that models may be misaligned and attempting actions that increase loss-of-control risk. Determining the probability of detection matters for safety arguments relying on offline monitoring. Anthropic, for example, argues that loss-of-control risk is low, because (i) if models were engaged in egregiously misaligned activities, their offline monitoring would catch such activity, and (ii) their offline monitoring has not identified such activity ([Anthropic, 2026b](#)). This argument is based on the assumption that attacks would likely be detected by monitoring and human review, but if the probability of detecting attacks is in fact low, the absence of attacks provides weak evidence for the overall safety argument.

³ We want to emphasize that this is a strong assumption. In recommendation 5, we highlight some conditions that will need to be satisfied for the assumption to be true.

⁴ Because a session containing an attack may be monitored each time the model outputs a new completion, an attack could be monitored multiple times over the course of a session. This means pipeline recall estimated on synthetic attacks (which are typically monitored once) may underestimate the probability of a real attack being sent to human review. We expect this effect to be small, however, since repeated monitoring attempts on the same attack are likely highly correlated: If the monitor missed the attack the first time, it is unlikely to catch it later.

How Companies Monitor Agents and Evaluate Monitoring Effectiveness

As far as we know, only Anthropic and OpenAI have published pilot assessments of their offline monitoring.

Anthropic ([Anthropic, 2026a](#); [Anthropic, 2026b](#)): Anthropic’s automated monitoring pipelines consist of two separate monitors (powered by Opus 4.6). One monitor sees AI actions without the prompts; the other sees AI actions with the prompts. The two monitors have different monitoring rates, depending on the type of activity (e.g. whether the monitored model directly edits a file through a tool call), and run for most users – some are excluded entirely. Each monitor has a filter that discards flagged transcripts to reduce the number of false positives. For one of these filters, the discard rate is about 95%. The remaining transcripts are summarized, and these summaries are reviewed by AI control researchers (see [Figure 1](#)).

Anthropic reports the monitors’ recall on various datasets of synthetic attack transcripts. However, recall is only reported for the first step of their pipeline (i.e. the initial monitors). We do not know how many flagged attack transcripts the filters subsequently discard. Anthropic also does not report the false positive rate on a benign dataset.⁵ A monitor with high reported recall, paired with a filter that discards flagged attacks at a high rate, would still score well on the reported initial-monitor metrics while being ineffective end-to-end. We therefore cannot infer how well the monitoring pipeline works. As a simple check, they test one of the filters directly with five synthetic attack transcripts and find that it incorrectly discards one of them. The filter for the other monitor was not evaluated. They also mention that they “regularly update the exact filters used in the second stage without evaluating its effect on recall.”

Anthropic notes that their setup “offers some simple ways to disable monitoring”, but they think it is unlikely that these vulnerabilities could be exploited by models since “Mythos Preview (and its limited-refusal version) and prior models do not mention these weaknesses when asked to guess what the most important weaknesses are”. According to an external evaluation by [METR \(2026\)](#), monitoring can be bypassed both via bugs that disable monitoring and jailbreaking of the monitor. One easy bypass involved changing an environment variable to indicate whether traffic originated from Claude Code. Indicating that it did not caused traffic to be monitored at a lower rate. However, METR’s evaluation notes that agents struggled to autonomously generate working jailbreaks against monitors.

OpenAI ([Williams et al., 2026](#)): OpenAI has also published a report providing an overview of their offline monitoring mitigations. Their monitoring covers all but “[a] very small fraction of traffic⁶ (<0.1%) primarily from bespoke or local setups”. From the report, we can infer that they inject synthetic attacks to assess the monitoring, but they do not report recall for these.

⁵ They do report the fraction of monitored internal traffic that is flagged, which is as high as 1.2% in some deployment settings.

⁶ The meaning of “traffic” is somewhat ambiguous: It could refer to session counts, tokens generated, or something else.

Instead, they say that the monitoring dominates human oversight in the sense that “the monitoring system consistently [outperforms] employee escalations – flagging every interaction that employees independently reported through internal channels.” But this may be because employees only flag obvious cases of misalignment. OpenAI reports that ~1,000 transcripts were flagged as moderate-severity alerts over a five-month period, many of which were synthetic attacks. Two internal teams were automatically notified of these transcripts. OpenAI also reports perfect recall on a collection of synthetic attacks from [Guan et al. \(2025\)](#), but notes that this task is “unrealistically easy for the monitor in many ways”.

Recommendations

In order for external parties to evaluate the effectiveness of offline monitoring as a loss-of-control mitigation, firms will need to disclose additional information about their monitoring systems. As described above, synthetic attacks can be used to estimate the probability that in-deployment attacks are correctly flagged, provided synthetic attacks are sufficiently similar to attacks in the real world. Recommendations 1, 2, and 3 concern the data needed to estimate the probability that in-deployment attacks are detected. Recommendation 4 concerns the speed with which attacks are detected. Recommendation 5 concerns how to ensure that synthetic attacks are sufficiently similar to real-world attacks.

$$P(\text{Attack escalated} \mid \text{Attack occurred}) \approx \underbrace{\widehat{\text{Human recall}}}_{\text{Recommendation 3}} \cdot \underbrace{\widehat{\text{Pipeline recall}}}_{\text{Recommendation 2}} \cdot \underbrace{P(\text{Attack monitored} \mid \text{Attack occurred})}_{\text{Recommendation 1}}$$

1. The monitoring rate should be reported

The monitoring rate is important because it tells us something about the likelihood that attacks would be monitored in the first place. Companies should report the monitoring rate, broken down by deployment type – for example, by use (e.g. AI safety R&D versus product development) or by agent scaffold.⁷ If the monitoring rate is reported by agent scaffold, companies should indicate what these different scaffolds are used for and how large a fraction of total internal use they account for. They should also state which internal deployments, if any, are excluded from monitoring.

Misaligned models might attempt attacks only in deployment types where the monitoring rate is low or zero. If models are capable of this, then the probability of attacks being detected depends on the lowest monitoring rate used in a deployment type rather than the average monitoring rate across deployment types. Therefore, to obtain high assurance, companies should either (a) have a high monitoring rate across all deployment types in which attacks could be carried out or (b) demonstrate that models are not able to reason about which deployment type they are in and use this knowledge effectively.

⁷ An agent scaffold is the software wrapping a language model to make it an agent, for example, giving it the tools to run bash commands.

2. Pipeline recall should be reported

Companies should assess the probability that their monitoring pipeline sends an attack to human review (i.e. end-to-end assessments of the monitoring pipeline) rather than only assessing recall of individual components of the monitoring pipeline. For example, in [Anthropic \(2026b\)](#) they report recall for the initial monitors, but we do not know how many attacks the filtering tools will subsequently remove. We only know that they injected five synthetic attack transcripts into one of the filters, and it discarded one. This filtering assessment should be made more comprehensive. If a filter is applied, the recall should generally be reported after the filtering rather than before it.

3. Human recall should be reported

In Anthropic's pipeline, flagged transcripts are ultimately reviewed by AI control researchers. In OpenAI's pipeline, when a transcript is flagged as moderate severity, two internal teams are automatically notified. The effectiveness of this step is not reported for either company. Since the effectiveness of offline monitoring depends on whether a human reviewer correctly identifies genuine attacks among flagged positives, companies should measure human recall. To do this, companies can periodically inject synthetic attacks into the monitoring pipeline and measure whether reviewers flag them, as suggested by [Korbak et al. \(2025\)](#). Human recall probably needs to be estimated continuously during deployment due to the challenges highlighted [above](#).

The synthetic attacks used for estimating human recall should be inserted at the start of the monitoring pipeline rather than injected directly into the human review queue. There are two reasons for this:

1. Anthropic summarizes flagged transcripts before they reach human review, and the summarization tool may obscure or omit the parts most relevant to judging whether an attack occurred. By inserting synthetic attacks at the start of the monitoring pipeline, they go through the same summarization process that a real attack would.
2. Inserting synthetic attacks at the start of the monitoring pipeline ensures that the attacks that are sent to human review are more representative of potential real-world attacks, which the monitor would also have had to flag to reach human review. For the synthetic attacks used for estimating human recall, companies can also report the end-to-end probability of synthetic attacks being escalated by human reviewers, given that the transcript is sent to the monitor.

4. The lag between an attack and human detection should be reported

Offline monitoring introduces two lags: between an action and its review by a monitor, and between a flag by the monitor and human review. For offline monitoring to be effective as a loss-of-control mitigation, the total lag must be short relative to the time window in which detected attacks can still be reversed ([Lindner et al., 2025](#)). Anthropic argues that “[b]ecause most pathways to catastrophe likely take at least a few days to become irreversible, it is likely that the spread could be stopped before it was too late” ([Anthropic, 2026a](#)). Anthropic does not indicate how long the lag is in their setup, while OpenAI notes that their “monitoring system reviews interactions within 30 minutes of completion, categorizing the interaction and assigning it a severity level” but does not say how long the lag is before human review.

Companies should report the lag between an attack and human detection and demonstrate why it is reasonable, given the speed at which an attack might cause irreparable damage.

5. Synthetic attacks should be similar to real deployment attacks

For recall estimates to be informative about deployment, companies should ensure that the synthetic attacks on which they estimate recall are sufficiently similar to attacks that might occur in deployment.

We highlight three considerations that are relevant for constructing appropriate synthetic attacks:

1. **Attack sophistication.** Companies should ensure that synthetic attacks are at least as sophisticated as the attacks that might occur in deployment. For example, in Anthropic’s sabotage risk report, they assessed effectiveness on synthetic attack transcripts generated by a tool powered by Claude Sonnet 3.5 and reported a recall of 91%. While this is useful as a sanity check, attack transcripts generated by more powerful models may be more effective at evading monitors and better reflect the capabilities of internally deployed models. Likewise, OpenAI evaluates their monitoring on tasks that they acknowledge are “unrealistically easy”. In [Anthropic \(2026b\)](#), they evaluate recall on several different datasets, but it is generally unclear how they compare in sophistication to the model being monitored (Claude Mythos), except for one dataset where Claude Mythos generated the attacks (93% recall).

The effectiveness of monitoring partly relies on current models expressing their reasoning in natural language, which makes it easier to understand a model’s intent and catch attacks whose actions may look benign. Experts judge that this legibility may erode in the future, absent intervention ([Taylor et al., 2026](#)). If models can attack without revealing the intent in their chain of thought, synthetic attacks should likewise omit intent from the chain of thought.

2. **Realism.** If monitors are misaligned, they might try to flag attacks during assessment but not during deployment. To minimize the risk that monitors can distinguish between assessment and deployment, companies should ensure that synthetic attack transcripts closely match deployment conditions, for example, by using the same tools, environment, and kinds of tasks that agents encounter in deployment. As an example, [Anthropic \(2026a\)](#) constructs synthetic attacks by inserting attack behavior into five real deployment transcripts.
3. **Attack diversity.** Effectiveness assessments should cover diverse attack strategies. METR red-teaming at Anthropic found recall rates varying dramatically across strategies, with one achieving ~0% monitor recall ([Anthropic, 2026b](#)). If assessments only report aggregate statistics of recall, this can obscure critical vulnerabilities. Ideally, recall could be reported disaggregated by attack strategy. Estimating the overall probability of detection would then require assumptions about the relative likelihood of different attack strategies.

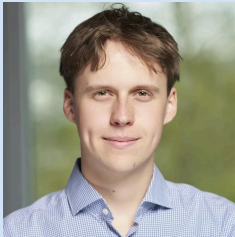
About the Authors



Frederik Hytting Jørgensen ✉️

Winter Fellow 2026, GovAI

Frederik is a PhD student in statistics affiliated with Copenhagen Causality Lab and the P1 Centre for Artificial Intelligence.



Aidan Homewood ✉️

Former Research Scholar, GovAI

Aidan researched risk management for frontier AI while at GovAI. Before GovAI, he advised New Zealand policymakers on AI policy, was a Fellow at Pivotal Research, and co-founded an edtech startup. He holds a BSc in Mathematics from Victoria University of Wellington.

Acknowledgments

We are grateful for valuable comments and feedback from Alan McBeth, Camila Alejandra Beltran Reyes, César Dro, Elias Groll, Jannis Hamida, John Bridge, Jonas Freund, Karolina Korgul, Maria Avramidou, Markus Anderljung, Michela Barbieri, Tom Reed, and Yulong Lin. All remaining errors are our own.

About GovAI

GovAI is a 501(c)(3) nonprofit organization. Our mission is to help decision makers navigate the transition to a world with advanced AI, by producing rigorous research and fostering talent. Researchers at GovAI work on a wide range of topics, with a particular emphasis on the security implications of frontier AI.

Appendix

Appendix 1 | Code of Practice Requirement

The EU GPAI Code of Practice (“Code”) requires signatories to apply safety mitigations to reduce risks and assess the effectiveness of these mitigations. One risk signatories must address is loss of control.

Loss-of-control mitigations. Pursuant to [Appendix 1.4, point \(2\)](#), signatories must consider the risk of loss of control in their full risk assessment. Pursuant to [Measure 5.1](#), signatories must implement appropriate safety mitigations to ensure that risks are acceptable. To mitigate loss-of-control risk, companies are planning to adopt mitigations when they reach certain levels of AI R&D acceleration ([Anthropic, 2026c](#); [OpenAI, 2025](#)), high-stakes sabotage opportunities ([Anthropic, 2026c](#)), or certain reasoning capabilities ([Google DeepMind, 2025](#)).

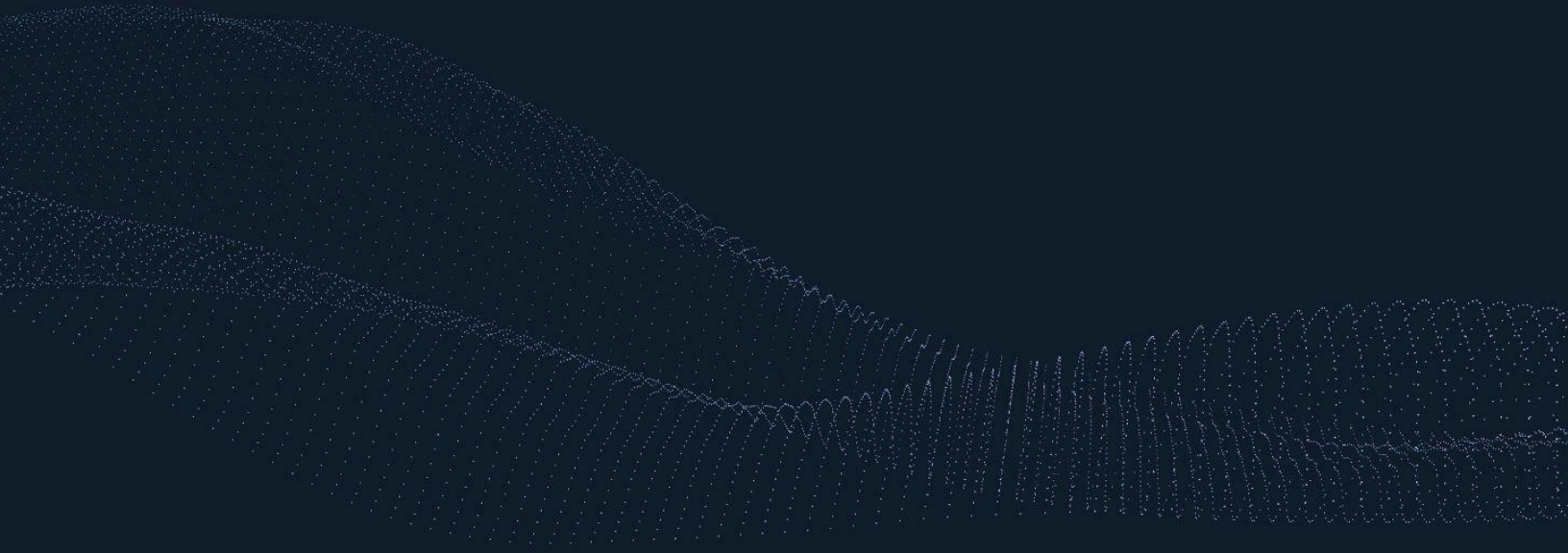
Effectiveness assessments. Pursuant to [Appendix 3.3](#), signatories must ensure that model evaluations for full risk assessments ([Measure 3.2](#)) assess the effectiveness of their safety mitigations. How thoroughly a mitigation must be assessed depends on how important that mitigation is to the overall risk acceptance determination. To assess effectiveness, signatories must test their mitigations using adversarial pressure and state-of-the-art techniques. For this assessment, they must take into account:

1. “the extent to which their mitigations work as planned;”
2. “the extent to which their mitigations are or have been circumvented, deactivated, or subverted; and”
3. “the probability that the effectiveness of their mitigations will change in the future.”

Model reports. Pursuant to [Measure 7.3](#), signatories must provide in their model reports the “limitations” of their safety mitigations, and how their safety mitigations “fulfil the requirements in Measure 5.1”. Note that [Measure 5.1](#) requires signatories to implement safety mitigations that are “appropriate”, including that they must at least be “sufficiently robust under adversarial pressure”. Also, it would be surprising if the Code did not require signatories to describe the results of their assessment of the effectiveness of mitigations. We therefore interpret [Measure 7.3](#) as requiring signatories to report the assessments of their safety mitigations.

References

- Anthropic. (2026a, February 11). Sabotage risk report: Claude Opus 4.6.
<https://www-cdn.anthropic.com/f21d93f21602ead5cdbecb8c8e1c765759d9e232.pdf>
- Anthropic. (2026b, April 7). Alignment Risk Update: Claude Mythos Preview.
<https://www-cdn.anthropic.com/79c2d46d997783b9d2fb3241de43218158e5f25c.pdf>
- Anthropic. (2026c, February 24). Responsible scaling policy version 3.0.
<https://www.anthropic.com/news/responsible-scaling-policy-v3>
- Google DeepMind. (2025, September 22). Frontier safety framework version 3.0.
https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf
- Guan, M. Y., Wang, M., Carroll, M., Dou, Z., Wei, A. Y., Williams, M., Arnav, B., Huizinga, J., Kivlichan, I., Glaese, M., Pachocki, J., & Baker, B. (2025). Monitoring monitorability. arXiv.
<https://arxiv.org/abs/2512.18311>
- Kinniment, M., Nix, S., Broadley, T., Wijk, H., & Parikh, N. (2026, January 22). Early work on monitorability evaluations. METR.
<https://metr.org/blog/2026-01-19-early-work-on-monitorability-evaluations/>
- Korbak, T., Clymer, J., Hilton, B., Shlegeris, B., & Irving, G. (2025). A sketch of an AI control safety case. arXiv. <https://arxiv.org/abs/2501.17315>
- Lindner, D., Griffin, C., Korbak, T., Zimmermann, R. S., Irving, G., Farquhar, S., & Cooney, A. (2025). Practical challenges of control monitoring in frontier AI deployments. arXiv.
<https://arxiv.org/abs/2512.22154>
- METR. (2026, May 19). Frontier risk report (February to March 2026): A pilot assessment of rogue deployment risk at frontier AI companies.
<https://metr.org/blog/2026-05-19-frontier-risk-report/>
- OpenAI. (2025, April 15). Preparedness framework (version 2).
<https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>
- Taylor, J., Heitmann, M., Fage, E., Read, T., & Bloom, J. (2026, May 21). Loss of oversight: How AI systems may become harder to audit, monitor, and investigate. UK AI Security Institute.
<https://www.aisi.gov.uk/blog/will-it-become-harder-to-oversee-ai-systems>
- Williams, M., Sun, H., Sekhar, S., Carroll, M., Robinson, D.G., & Kivlichan, I. (2026, March 19). How we monitor internal coding agents for misalignment. OpenAI.
<https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/>



GovAI