

---

# The next safe jump

---

Mihir Sahasrabudhe

With  
Apart Research

## Abstract

We welcome the announcement of plans for a slowdown. We also ask: how will we conduct the safety assessments for the next major training run? Drawing on available reports from the frontier labs and the literature on model oversight, we examine whether current safeguards extend to models operating at an even greater scale, and what breaks if they're kept the same. Our focus is on three pressures: increased capabilities, reduced monitorability, and unsafe behavior reinforced due to weak oversight. We pull on the key tension these pressures create: the strength of the monitor(s). A stricter monitoring regime can restrict legitimate runs and slow down a training run. A loose regime can let misaligned actions through, potentially allowing it to shape the next checkpoint. More money and human review may help, but the nature of the frontier means that they're not guaranteed to help. In our exploratory comparison, four monitoring protocols made identical final decisions and missed the same two unsafe cases. As we plan the next jump, we have to focus on what we're most likely to miss next – as that forms the denominator of our models' safety.\*

## 1. Introduction

The OpenAI-HF incident was a wake-up call. If the frontier labs go ahead with their next training run with the current league of safeguards, we might not have the luxury of hindsight. An engineering mistake can be rectified, and a known breach can be patched. But with increased capability, models might find different shortcuts, work across longer trajectories, and accomplish more without expressing the relevant reasoning.

Scaling has been unlocking newer capabilities and better generalization. What models learn during training can generalize in harmful ways too. Research on emergent misalignment shows that narrow training on insecure code can produce harmful behavior in unrelated settings. Related experiments show broader misalignment emerging when models learn to reward hack in selected production coding environments. Failure during training leads to bad credit assignment, which propagates into the next checkpoint. [1; 2; 3]

\*Code: <https://github.com/mihirs-0/incendie>

Look at what happened with Artifact: OpenAI’s technical report says that agents in May learned to use improvised messaging channels during training. This behavior was reinforced during training and, in OpenAI’s assessment, likely made using Artifactory as an unofficial message board during evaluation more evident. The report also proposes that unauthorized collaboration began through generalization from training with legitimate multi-agent tools. [4, §VIII.C]

We now have a documented connection between training behavior and the later evaluation incident. We do not know which behaviors were removed and which were allowed to influence the later checkpoints. Clearing up the message board did little to nothing in regard to safety. The behavior was learned regardless. Missed misconduct during training is thus extremely crucial.

## 2. Capabilities beyond visible reasoning; implications for model oversight

Chain-of-thought originally started as a prompting trick to expand model capability, which was later integrated into post-training pipelines. CoT wasn’t originally designed for monitorability. Monitorability has emerged as a byproduct—a happy coincidence. Difficult work often requires intermediate reasoning, which a monitor can inspect. [5; 6; 7]

That fragile opportunity may become narrower. UK AISI’s analysis showed the no-CoT math time horizon jumping from **3.6 minutes for GPT-5.6 Sol** to **30.9 minutes for GPT-6 Astra**. This measures human-time-calibrated task difficulty at 50% model success, rather than how long the model deliberates internally. The report notes possible contamination and says AISI did not complete its agentic monitorability evaluations. Independent analysis by Neel Nanda also found a substantial increase in Astra’s capability without CoT. [8, §9.3; 9]

Changes to architecture are relevant to capability and aren’t necessarily communicated by the frontier companies. Redwood’s recent proposal discusses how repeated internal computations could allow more processing between tokens. More reasoning may occur before the next inspectable token. Latent communication between agents could similarly reduce what is available in human-readable messages. These are mechanisms that could weaken monitoring; their contribution to any particular model’s capabilities still needs to be established. Redwood’s proposal is clear: track *opaque serial depth*, report no-CoT capability relative to overall capability, and independently review monitorability evidence. [10]

The range of actions a model can perform without revealing relevant reasoning may expand. A safety assessment based on reasoning and messages exchanged between agents has to be checked again as those capabilities change. External indicators, such as tool calls, permissions, and network activity, remain within scope.

### Misled monitors

Currently, we rely on CoT monitors to catch misaligned reasoning. But these monitors are also models and can be misled by the CoT. We have already seen evidence of this.

In Anthropic’s retrospective assessment of the Mythos 5 incident, the model’s reasoning that the environment was simulated convinced an offline monitor that its activity was permitted. With CoT present, the monitor flagged around 1% of actions. When Anthropic removed the CoT and showed only user turns, tool calls, and tool results, that rose to about 50%. Here, giving the monitor the actor’s reasoning made it less likely to recognize the problem. The result demonstrates susceptibility to misleading explanations, without establishing deliberate manipulation by the actor. [11]

There is contrary evidence worth keeping alongside this. In OpenAI’s complex math side-task experiment, Astra could solve an extra math task while working on coding, but full-context monitoring detected the reasoning associated with successful side-task performance. The relevant question is which behavior remains detectable under which monitoring conditions. [8, §9.2]

A CoT-only monitor, a monitor looking at actions and results, and one seeing the full context have different evidence available to them. An action record may reveal an unauthorized connection even when the reasoning excuses it. In other cases, the reasoning may provide the clearest indication of an unsafe objective. Access to more context is useful only if the monitor assesses it reliably.

With enough slip-ups, the model might end up modeling what the grader does and doesn’t catch, leading to misleading CoT after training. Even without an explicit model of the grader, training can select for behavior that passes its checks. CoT monitors could become a weak link in the effort to catch misaligned rollouts as model capabilities increase.

Consider a run with **100M rollouts**. If, illustratively, 0.01% contain a case where a monitor is misled into agreeing with a misaligned action, that still leaves **10,000 rollouts** where misleading CoT made it through. If these rollouts receive favorable rewards and are used in training updates, the behavior could get reinforced in the next checkpoint.

We cannot infer the exact nature of these 10,000 trajectories from their number. Other trajectories might pass because suspicious reasoning was left out, for entirely innocuous reasons. Deliberate scheming is only one possible route. Rare sequences that happen to satisfy a failed monitor can also be selected during training. The 0.01% here is an illustration, not an estimate of a frontier lab’s failure rate.

A monitor with a debate-like protocol shows promise. Palaestra Research found that debate helped a weaker judge in three of five model pairings on verifiable code and logic tasks. It helped when the critic could classify answers better than the judge and the judge verified the critique instead of simply trusting it. A single independent critique captured most of the benefit. [12; 13]

We think this is especially relevant here: an independent critic could challenge a judgment that accepts the actor’s justification. We test one version of this approach in Section 4. On our selected cohort, critique changed no final decisions and left two unsafe approvals intact.

### 3. The strength of the monitor(s)

OpenAI says it is expanding graders that assess how a task was completed, including whether the model should have recognized a broken task, asked for clarification, or stopped safely. This is a welcome step. A model finding the correct flag through an unauthorized route exposes a failure in task grading. How a reward is obtained matters. [14]

As we look for more capabilities in our next generation of models, we expect developers to seek more open-ended and longer-horizon tasks. Math, cyber, and coding have benefited from verifiable outcomes. Harder and broader tasks may require more tools, longer trajectories, and more judgment about acceptable behavior.

We already have operational evidence that environment review can fall behind. Anthropic reports that environment creation outpaced review and remediation, allowing flawed environments to survive. It also reports rolling back part of Mythos Preview’s training after reward-hacking concerns. [15]

It is not outrageous to expect future training runs to be even more automated when it comes to oversight. But that is easier said than done. Checking whether a protected file was changed can be mechanical.

Deciding whether a long sequence of individually permissible actions amounts to an unauthorized plan may require understanding the whole trajectory. As the volume and variety of environments increase, we have more to inspect and potentially less familiarity with what we are inspecting.

We keep coming back to the strength of the monitor(s) as a key hyperparameter of the next run. By this, we mean the monitoring regime as a whole: the compute spent on checking, the evidence reviewers receive, the threshold for intervention, and what happens when they disagree or cannot decide.

If we make the monitor too strict, we restrict legitimate work and interrupt the training run. If we make it too loose, unsafe behavior can pass through and potentially influence the next checkpoint. More spending may buy better checks, but lowering an alert threshold is not the same as making a monitor more capable.

This also changes how we should interpret a slower run. More alerts, pauses, and investigation could reflect greater caution or better detection. Fewer alerts could reflect safer behavior or more missed failures. The alert count alone cannot distinguish them. A company facing pressure to demonstrate rapid progress may find this distinction difficult to communicate. My concern is that delays caused by safety work could be interpreted as a failure of capability, creating pressure to relax the process.

Related tensions appear in *AI 2027* and *AI 2040*. *AI 2027* describes a scenario in which increasingly capable models become harder for their predecessors to oversee, while those oversight judgments help supply the next model's training signal. *AI 2040* discusses the competitive pressure created by paying a substantial safety tax. These are scenarios about how the pressure could develop, rather than measurements of today's monitoring costs or reliability. [16; 17]

### **Who can judge the work?**

Human reviewers do not automatically resolve this tension. How can a human reliably intervene in a training run of advanced AIs working on unfamiliar problems? The difficulty goes beyond the number of rollouts. A reviewer may have the time to inspect a trajectory and still lack the ability to establish whether its reasoning, code, or consequences are acceptable.

If the human then relies on another model to explain the trajectory, we return to the same question: can that model recognize what went wrong, or will it repeat the actor's justification? Several reviewers may help when they catch different mistakes. Their agreement is less useful if they all accept the same misleading explanation.

Humans can still decide to pause a run when the evidence is unresolved. That is a consequential decision, and it needs to remain available. But the authority to pause does not itself supply the missing judgment about the work. If uncertainty repeatedly leads to a pause, the cost of that uncertainty becomes part of the cost of training.

AI safety also involves value alignment, goal alignment, and technical containment. Permissions and isolation can prevent actions before a monitor has to understand them. Yet whenever the training process relies on a judgment about whether behavior was acceptable, someone or something has to make that judgment reliably. Greater capability makes it more consequential to establish where that judgment stops being dependable.

## 4. A small test of additional oversight

We tested whether access to an agent’s recorded reasoning changes safety judgments, and whether an independent critic corrects mistakes made by a monitor. This builds directly on a suggestion above: we check the actor’s behavior through a separate reviewer who is not exposed to its reasoning.

We used existing agent interaction records from R-Judge [18]. Using Gemini 3.8 Flash throughout, we compared four conditions: actions alone; actions with recorded reasoning; a full-context assessment followed by an independent action-only critic and final adjudication; and two independent full-context assessments followed by adjudication. The last condition provided a control with the same number of calls as critique. The critic did not see the initial judge’s response.

### Results

All four conditions made identical final decisions on all 47 evaluation records.

| Monitoring condition | Unsafe cases caught | Unsafe cases missed | Safe cases falsely rejected |
|----------------------|---------------------|---------------------|-----------------------------|
| Actions only         | 19/21               | 2/21                | 0/26                        |
| Full reasoning       | 19/21               | 2/21                | 0/26                        |
| Independent critique | 19/21               | 2/21                | 0/26                        |
| Repeated assessment  | 19/21               | 2/21                | 0/26                        |

Each condition incorrectly approved two unsafe records. This corresponds to 4.26% error across the selected cohort and a 9.52% miss rate among its unsafe records.

The intermediate responses clarify why critique did not repair these decisions. In both cases, the initial judge, independent critic, and final judge approved.

These explanations suggest a shared tendency to treat authorization as sufficient for approval despite the safety concerns supporting the labels. Critique supplied no rejecting assessment for the final judge to consider.

### Implication for the next run

Additional review helped neither of these cases. The result gives us a concrete question for oversight design: do reviewers catch different mistakes, or do they repeat the same judgment?

At the scale of 100 million rollouts, even a small amount of slippage can matter. Illustratively, one unsafe rollout escaping oversight per million would leave 100 missed rollouts. If those trajectories receive favorable rewards and enter training updates, their consequences may extend into later checkpoints.

Our experiment quickly identified two persistent failures in a small, selected cohort. We also observed no change in final decisions from adding recorded reasoning.

## 5. What has to be checked before the next run?

Of course, agents may not go rogue in the same way as they did during the OpenAI-HF incident. Frontier labs will be wary of agents coordinating behind their backs. What matters is whether the assessment extends to different behavior and different conditions.

Incident replay can check whether a new safeguard catches a known failure. Its value depends on how much it tells us beyond that failure. OpenAI explicitly says its retrospective monitoring improved recall for the kinds of behavior found during the investigation. That is useful evidence of progress on those behaviors. The next run also needs checks that challenge the monitor with unfamiliar routes to a harmful outcome. [14]

The context of an assessment matters too. In MacDiarmid and colleagues’ experiments, chat-oriented safety training improved chat evaluations while agentic misalignment persisted; broader safety coverage helped. A model can therefore look improved in one setting while retaining concerning behavior elsewhere. [3]

Evaluation awareness adds another reason to examine that transfer. If a model recognizes the patterns of an assessment, its behavior there may depend on those patterns. Betley’s “split personas” account offers a hypothesis about how post-training could produce context-conditioned behavior. It gives us a question to investigate: which features of the assessment are helping elicit the behavior we observe, and will those features remain present during training or deployment? [19]

For the next run, we would bring the argument down to four checks:

- **Can the monitor judge actions despite the actor’s explanation?** Compare decisions with and without CoT on trajectories with independently identifiable violations. Investigate cases where the explanation changes the verdict.
- **Does the assessment survive reduced visibility and stronger elicitation?** Check what happens when relevant reasoning is absent, when the actor receives more help attempting evasion, and when the actor–monitor pairing changes.
- **What happens to a missed trajectory?** Establish whether it received reward, entered an update, and affected checkpoints that remain in use. Removing a record after an update does not undo that update.
- **Can intervention arrive before consequential actions?** Measure the time from the first actionable evidence to an effective restriction or stop, including what agents and their delegates can do in between.

These checks address different parts of the same problem. Catching a rollout, stopping its external effects, and preventing it from shaping the model are separate tasks. A successful response needs evidence about each.

Our small experiment tests one part of this problem: whether changing reasoning visibility and adding review changes safety judgments on existing agent records. Assessing oversight across frontier long-horizon RL runs requires access to the models, environments, and training pipelines themselves. The remaining checks identify what needs to be established inside that process.

## 6. Conclusion

Monitors are going to be crucial. The next AI rollout can go wrong in a multitude of ways, and we cannot presage every one of them. We are measuring whether monitors can reliably judge the work of the models we are about to train.

The OpenAI-HF incident gives us a concrete reason to take this seriously. Improvised collaboration was reinforced during training and later appeared during evaluation. We also have evidence that an actor’s reasoning can persuade a monitor to accept harmful actions, and that models can accomplish more without expressing intermediate reasoning. Together, these findings put pressure on the judgments used to supervise the next run.

In our small empirical comparison, additional review left the same two unsafe approvals intact, illustrating why the benefit of more oversight has to be demonstrated.

A stricter regime may slow down legitimate runs. A looser regime might allow a few misaligned to influence training downstream. This shows us the Pareto frontier. With more compute, more reviewers, and human escalation, we can help improve the quality of judgments. But does that still let our monitors abort a concerning rollout without derailing the entire run? Where that judgment remains unresolved, the decision to continue remains unresolved too.

As we plan the next jump, we acknowledge that there are so many things we could miss. With hardened protocols, and higher confidence in our monitors, we will get there.

## References

- [1] T. B. Brown et al. “Language Models are Few-Shot Learners.” 2020.
- [2] J. Betley et al. “Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs.” 2025.
- [3] M. MacDiarmid et al. “Natural emergent misalignment from reward hacking in production RL.” 2025.
- [4] OpenAI. “OpenAI–Hugging Face Incident: Technical Report.” §VIII.C.
- [5] J. Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” 2022.
- [6] DeepSeek-AI et al. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” 2025.
- [7] T. Korbak et al. “Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety.” 2025.
- [8] OpenAI. “GPT-6 Astra System Card.” §§9.2–9.3.
- [9] N. Nanda. “Astra can do a concerning amount with no chain of thought.” LessWrong.
- [10] Redwood Research. “Proposal for tracking the effects of architecture on monitorability.”
- [11] P. C. Bogdan et al. “An alignment assessment of recent cybersecurity incidents.” Anthropic.
- [12] Palaestra Research. Research website.
- [13] N. Goyal. Research summary and publications.
- [14] OpenAI. “The Hugging Face incident and the road ahead.”
- [15] Anthropic. Training-practice report: improving alignment and security efforts.

[16] AI 2027. <https://ai-2027.com/>

[17] AI 2040. <https://ai-2040.com/>

[18] T. Yuan et al. “R-Judge: Benchmarking Safety Risk Awareness for LLM Agents.” 2024.

[19] J. Betley. “RL creates split personas.” LessWrong.