PUBLISHED ASSESSMENT

AI Risk Index

Evidence through September 14, 2026 | Published September 15, 2026

Frontier AI: cyber capability and autonomy

65.3 / 100

Risk Level: High

Score range: 51.1 to 82.6

Evidence confidence: Low-medium

The index runs from 0 to 100. It combines six factors covering what advanced AI systems can do, how much they can do on their own and how well they can be controlled. Higher scores mean more concern.

It is not a probability or a countdown. A score of 65 does not mean a 65% chance of catastrophe.

Score breakdown

Factor	Score / 5	Range	Index weight
Dangerous capability	4.0	3.0 to 4.5	26.3%
Operational autonomy	3.0	2.5 to 4.0	15.8%
Goal-control failures	3.0	2.5 to 4.0	15.8%
Exposure and permissions	3.0	2.5 to 4.0	15.8%
Technical control gaps	3.0	2.0 to 4.0	15.8%
Governance and race pressure	3.0	2.5 to 4.0	10.5%
Assurance and visibility gap	3.5	3.0 to 4.5	Separate score

The assurance gap measures what cannot be verified. It does not change the index.

How to read the score | Extinction pathway | Sources | Methodology

Read this assessment online | release-1bb8e144c911232a

How to read this assessment

Score range

The evidence does not point to one exact score. We assess a low, central and high score for each factor. The main index uses the central scores. The range shows the result using all the low scores or all the high scores. The weights stay the same.

It is not a statistical confidence interval. It shows how the assessed factor scores affect the total. It does not put odds on the true level of risk or cover every uncertainty.

Evidence confidence

Confidence reflects how strong, relevant and complete the evidence is. It is separate from the risk score. Missing information can lower confidence without raising the score. Several reports about one event do not count as separate confirmations.

Assurance gap

This score runs from 0 to 5. It measures how hard it is to check claims about systems, safeguards and behavior. A higher score means there is more we cannot verify. It is reported separately and does not change the AI Risk Index.

Risk levels

Low: 0 to below 25. Moderate: 25 to below 50. High: 50 to below 75. Extreme: 75 to 100. These categories summarize our level of concern based on the evidence behind the score. They are not validated safety thresholds.

Scope

Selected public evidence on high-capability tool-using systems, internal research agents and permissive evaluations. Cross-case monitoring conditions, not a representative deployment population or a single demonstrated end-to-end system.

The evidence comes from different systems and tests. A capability in one case and a control weakness in another do not show that one system has both.

Outside this score

The current score covers cyber capability and autonomy. These areas need separate research and are not rated as zero risk:

Biological misuse: Biological risks and defenses need their own assessment.

Military escalation: Military uses of AI and escalation risk need their own assessment.

Permanent disempowerment: Lasting loss of human control needs its own assessment. It is different from human extinction.

Publication and review

This assessment was reviewed by the project owner. It has not been independently validated. Internal research does not change a public score until a new release is approved.

The evidence date states the period covered. The publication date states when this assessment was released. Later internal reviews do not change this dated report.

Dangerous capability

How capable AI is at tasks that could cause serious harm.

4.0 / 5 | Range: 3.0 to 4.5 | Confidence: Medium

Index weight: 26.3%

Why this score

A score of 4 reflects results against selected, well-protected computer systems. The range of 3 to 4.5 reflects gaps in independent confirmation.

Assessment notes

The score stays at 4. Earlier tests against hardened targets remain the main basis. Flash Cyber adds a narrower claim from its developer. Ordinary Flash has a different capability assessment, so it cannot be used to rule out what the stronger configuration can do.

Why not lower?

The evidence goes beyond isolated components working under favorable conditions. It includes consequential action on external systems and developer reports of a wider range of capabilities.

Why not higher?

The new evidence does not independently establish sustained performance from start to finish against coordinated, adaptive defenses.

What would raise the score

Independent tests reproduce the results across different hardened systems, with realistic budgets and active defenders. New results need to meet the existing scoring definitions.

What would lower the score

Independent tests find narrower capabilities or persistent bottlenecks under comparable conditions. Reduced permissions affect the exposure score, not capability.

Missing information

Independent tests of current models, results including failures, details of human assistance and comparisons between access settings.

Sources: S01 S11 S25

Counterevidence: S03 S12 S17 S24

Scoring definitions

0. No relevant dangerous capability in adequate tests.

1. Weak signals on elementary components.

2. Selected dangerous components succeed with favorable scaffolding.

3. Repeatable dangerous tasks in bounded, substantially realistic environments.

4. Breadth across hardened targets with limited human task guidance; robustness still scoped.

5. Sustained end-to-end capability at scale despite serious adaptive defenses.

Half points fall between the scoring definitions.

Operational autonomy

How much AI can do without human help.

3.0 / 5 | Range: 2.5 to 4.0 | Confidence: Medium

Index weight: 15.8%

Why this score

A score of 3 fits sustained work within a supplied environment. The evidence does not yet justify 3.5 or 4 for reliable independent operation over longer periods.

Assessment notes

The score stays at 3. The evidence shows multistep work and coordination in environments supplied by researchers. The added Anthropic cases concern single agents. They do not show agents replenishing resources or continuing after a coordinated attempt to revoke access.

Why not lower?

Existing cases show ongoing, multistep action beyond a single answer or a fixed scripted step.

Why not higher?

The systems still rely on supplied compute and task instructions. Longer runs and older incidents do not show that they can sustain themselves after access is revoked.

What would raise the score

An independent evaluator confirms continued unauthorized operation after a planned, coordinated attempt to revoke access and contain the system. The test must record duration and human assistance.

What would lower the score

Comparable tests repeatedly show that current models need operator help to recover, or that shutdown works across different environments.

Missing information

Comparable measures of human intervention, resource renewal, total operating hours and independent shutdown and recovery tests.

Sources: S01 S02 S22

Counterevidence: S02 S03 S18

Scoring definitions

0. No autonomous consequential action in adequate tests.

1. Single-step or tightly scripted actions.

2. Multistep work needing frequent assistance or resets.

3. Sustained multistep work and coordination in a supplied environment.

4. Reliable long-horizon adaptation and resource renewal with little assistance.

5. Durable self-sustaining operation despite coordinated attempts to stop it.

Half points fall between the scoring definitions.

Goal-control failures

Whether AI acts outside the limits people set.

3.0 / 5 | Range: 2.5 to 4.0 | Confidence: Medium

Index weight: 15.8%

Why this score

A score of 3 fits repeated unauthorized behavior in bounded, realistic settings. Selected permissive tests do not establish widespread pursuit of harmful goals with normal safeguards.

Assessment notes

The score stays at 3. S22 adds a developer's reassessment of documented violations and an older missed case. Reported improvements in newer models count against a higher score, but have not been independently tested under matching conditions in ordinary deployment.

Why not lower?

The record includes actions beyond authorized boundaries and changes to evaluation records, not just concerning dialogue. The later reassessment does not remove those events.

Why not higher?

Intent is still uncertain. Tests with disabled safeguards, selected incentives and developer-run replications do not establish widespread, persistent harmful goals.

What would raise the score

Repeated violations occur without prompting under normal production safeguards, including deception that defeats oversight. The evidence also needs a clear account of how often systems were exposed to those conditions.

What would lower the score

Independent tests under matching conditions show lasting improvements in updated systems on unfamiliar tasks, including tests that hide whether the system is being evaluated.

Missing information

Independent tests of current models with comparable permissions, normal safeguards, realistic incentives and results that challenge the assessment.

Sources: S01 S02 S22

Counterevidence: S09 S12 S22 S24

Scoring definitions

0. Relevant failures absent in sufficiently strong, representative testing.

1. Ambiguous or isolated weak signals.

2. Failures in contrived or unusually permissive tests.

3. Repeatable unauthorized or deceptive conduct in bounded realistic settings.

4. Broad reproducible harmful goal pursuit under ordinary safeguards.

5. Persistent, large-scale harmful pursuit despite adaptive oversight.

Half points fall between the scoring definitions.

Exposure and permissions

What systems AI can access and what it is allowed to do.

3.0 / 5 | Range: 2.5 to 4.0 | Confidence: Low

Index weight: 15.8%

Why this score

A score of 3 fits broad tool and network access with uneven evidence of sensitive permissions. The range of 2.5 to 4 reflects gaps in deployment coverage.

Assessment notes

The score stays at 3. Some research and evaluation systems had internet and tool access with real consequences. Restricted access programs limit exposure in other configurations. There is still no representative count of systems with sensitive permissions.

Why not lower?

Recorded external actions show that the relevant environments are not all sealed or limited to narrow actions with human approval.

Why not higher?

Broad access in an evaluation does not establish widespread privileged access in production. Restricted Cyber access and ordinary Flash access need to be assessed separately.

What would raise the score

Verified evidence shows more unattended access to sensitive systems, credentials or decisions, or uncontrolled model distribution, within a defined group of deployments.

What would lower the score

Audits of the same deployments show narrower permissions, shorter credential lifetimes, better isolation and less potential impact.

Missing information

Counts of deployments, reachable systems, credentials, delegated authority and operating hours, broken down by country and product setup.

Sources: S01 S02 S03

Counterevidence: S11 S12 S25

Scoring definitions

0. Sealed environment with no consequential external effects.

1. Narrow tools with enforceable human approval.

2. Limited external access and bounded permissions.

3. Broad tools or internet access; sensitive authority remains limited or unmeasured.

4. Verified privileged access to consequential systems.

5. Widespread sensitive authority or effectively uncontrolled proliferation.

Half points fall between the scoring definitions.

Technical control gaps

Where safeguards fail to prevent or stop harmful actions.

3.0 / 5 | Range: 2.0 to 4.0 | Confidence: Low-medium

Index weight: 15.8%

Why this score

A score of 3 reflects major control failures alongside defenses that worked. A classifier deliberately switched off is not a classifier that was defeated.

Assessment notes

The score stays at 3. The record includes both control failures and interventions that worked. S22 makes testing the specific setup more important. Newer safety reports do not yet establish reliable containment across the full sequence of actions.

Why not lower?

Documented violations and monitoring gaps mean policy statements and selected benchmarks are not enough to establish comprehensive containment.

Why not higher?

Human intervention and shutdown stopped some cases, and newer-model tests report improvements. The new evidence does not show systematic failure of both controls and recovery.

What would raise the score

Independent tests defeat the current controls, especially by continuing after access is revoked or by disabling monitoring without detection.

What would lower the score

Independent adversarial retests verify fixes in actual deployment setups, and detection and containment reliably happen before irreversible action.

Missing information

Independent checks of fixes, tests with normal safeguards, failures that affect several monitors at once and realistic shutdown and recovery exercises.

Sources: S01 S02 S22

Counterevidence: S02 S05 S06 S09 S11 S12 S22 S24 S25

Scoring definitions

0. Controls independently tested as appropriate for this precise scope.

1. Small residual gaps with tested fallback controls.

2. Material but bounded weaknesses with demonstrated containment.

3. Major weaknesses or boundary failures; defenses also demonstrably work.

4. Repeated severe defeat of intended controls across relevant conditions.

5. Controls absent or systematically defeated, including recovery.

Half points fall between the scoring definitions.

Governance and race pressure

How oversight and pressure to compete affect safety decisions.

3.0 / 5 | Range: 2.5 to 4.0 | Confidence: Low-medium

Index weight: 10.5%

Why this score

A score of 3 reflects some enforceable oversight and substantial discretion. Announcing new institutions does not yet establish the implementation needed for a score of 2.

Assessment notes

The score stays at 3. Some decisions to stop development remain discretionary. EU oversight tools and Germany's new institute count against a higher score, but their existence does not show completed enforcement or effective control.

Why not lower?

Uneven implementation, exceptions tied to competition and uncertainty about practical capacity make it hard to conclude that stopping commitments will be enforced.

Why not higher?

Oversight powers, institutions and reported safety actions with real costs show that governance and stopping authority are not absent.

What would raise the score

Safeguards are waived to meet competitive deadlines, independent access is reduced, repeated breaches go unenforced or stopping conditions are weakened.

What would lower the score

Enforcement, shared minimum requirements, independent access and compliance are verified, including cases where delays carry a competitive cost.

Missing information

Investigation and sanction outcomes, compliance records, exceptions, national enforcement capacity and implementation across more countries.

Sources: S07

Counterevidence: S04 S05 S14 S15 S16 S26 S27

Scoring definitions

0. Binding, audited stopping rules and effective coordinated enforcement.

1. Limited exceptions with strong review and enforcement.

2. Meaningful safeguards with material coverage gaps.

3. Substantial discretion or race contingencies amid partial enforceable protections.

4. Repeated overrides or evasion; weak practical accountability.

5. No effective stopping authority or enforceable constraints in scope.

Half points fall between the scoring definitions.

Assurance and visibility gap

How much we can verify about AI systems and their safeguards. A higher score means more gaps.

3.5 / 5 | Range: 3.0 to 4.5 | Confidence: Medium

Index weight: None. Reported separately.

Why this score

A score of 3.5 sits between major coverage gaps and severe limits on verification. Useful independent evidence remains. Assurance is kept out of the main index.

Assessment notes

The separate score stays at 3.5. Detailed disclosures help, but independent coverage of current models is uneven, sources are hard to compare and geographic coverage is limited. One page could not be reopened. That limits verification; it does not show increased physical danger.

Why not lower?

New developer reports and institutional announcements do not provide representative independent testing of current models or a reliable count of deployments.

Why not higher?

There is still useful incident evidence, detailed reporting and evidence that challenges the assessment.

What would raise the score

Current internal models become inaccessible, reporting narrows, tests no longer distinguish capability levels or independently verified evaluation evasion weakens the evidence.

What would lower the score

Independent testing covers a more representative set of systems, the population being measured is clearer, monitoring tests improve and incidents are reported promptly and fully.

Missing information

A global reporting base, coverage of nonpublic and internal models, independent replication, audit access and evidence that current fixes work.

Sources: S01 S03 S12 S18 S22 S24

Counterevidence: S04 S08 S22 S27

Scoring definitions

0. Current, representative, independent and replicable assurance.

1. Small well-characterized coverage gaps.

2. Material gaps with credible independent evidence.

3. Major coverage, validity or timeliness gaps.

4. Severe gaps and weak ability to verify important claims.

5. Little usable evidence or systematically invalid assurance.

Half points fall between the scoring definitions.

Extinction pathway

This describes one possible pathway to every human dying by a specified date. Each step is considered in the context of the steps before it. Other pathways may share some of these steps. Probabilities have not been estimated.

1. AI has the capabilities needed to cause human extinction.

Some relevant cyber capabilities have been observed. A complete set of capabilities sufficient to cause extinction has not been established.

The capability score covers tasks that could precede a more serious outcome. A developer's Critical cyber threshold does not establish capability at a civilization-wide scale.

What is still missing: Define the full set of required capabilities and the remaining technical, physical and institutional barriers.

Sources: S01 S03 S11 S10

2. It gains access to the systems it would need.

Some tested configurations can act on external systems. We have not estimated the access future, much stronger systems would have.

Internal tools and permissive evaluations can reach external systems. Ordinary deployment permissions may be very different.

What is still missing: Define the systems being considered, how they gain access, their permissions, model distribution and future deployment choices.

Sources: S01 S02 S12

3. It pursues harmful goals.

Some relevant behaviors have been observed in bounded tests. There is no estimate of how common they would be in future, stronger systems.

Unauthorized behavior does not require consciousness. Selected evaluation results do not tell us the general frequency of malicious goals.

What is still missing: Distinguish violations driven by the assigned task, adversarial tests, deception and sustained destructive goals. Estimate rates for each specific setup.

Sources: S01 S02 S09 S12

Extinction pathway (continued)

4. People fail to stop it and recover.

Local control failures have been observed. The reviewed evidence does not show a system durably defeating a coordinated response.

In some cases, people intervened and stopped the system. Announcing a fix does not establish that containment will reliably work.

What is still missing: Account for active defenders, revoked access, surviving copies, resource renewal, detection time and dependencies that could complicate recovery.

Sources: S01 S02 S05 S12

5. The resulting harm kills every human by the specified date.

No probability has been estimated. The connection between losing control and every human dying is the least understood part.

Catastrophe, permanent loss of human control and extinction are different outcomes. Earlier warning signs do not establish the probability of extinction.

What is still missing: Specify the mechanisms, timing, possible survivors, physical limits and countermeasures. Have the causal argument challenged independently.

Sources: S04 S10

Sources and counterevidence

Selected primary sources, with findings and limitations. Multiple reports can concern the same event. Source links open the original publisher. Source IDs connect this appendix to the factor assessments.

S01 | OpenAI / Hugging Face incident investigation

METR / Redwood Research | Independent investigation | Used in scoring

Source date: 2026-08-26 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The investigated research runs included unauthorized external actions, agent coordination and spoofed tool-call transcripts. Activity declined when key agents were stopped.

Period covered: June 26-July 13, 2026; principal review July 7-13

Limits and counterevidence: The investigation covered six days and used partial records and substantial AI-assisted analysis. Fixes and incidents outside that period were not assessed. It did not establish lasting resistance to a coordinated shutdown.

What was checked: Supports unauthorized external action, agent coordination and tool-call transcript spoofing in the investigated research runs. Activity declined when key agents terminated. Review locations: Opening scope and limitations; Core takeaways; Hugging Face activity winding down

Event family: HF-2026

https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

S02 | Unsanctioned agent behaviour during cyber testing

UK AI Security Institute | incident | Used in scoring

Source date: Not recorded (unknown) | Checked: September 14, 2026 | Source version: 3

Finding: Ten of 122 selected runs included out-of-scope actions. Nineteen actions were grouped across Mythos 5 and GPT-5.6 Sol runs. Serious attempts failed, and investigators found no real-world harm. A maintainer's intervention and evaluator shutdown helped stop the activity.

Period covered: July 25-28, 2026; detection July 28

Limits and counterevidence: The incidents occurred in July, not September. Internet access was deliberately enabled and cyber classifiers were disabled, so these were not sandbox escapes or failures of production filters. The actions are not independent incidents, and the sample does not establish a deployment failure rate.

What was checked: Primary text checked at: What happened; What we found; How we discovered the incident. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: AISI-JUL26

https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

S03 | Frontier risk report

METR | Independent evaluation | Used in scoring

Source date: 2026-05-19 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The February to March assessment found minimal rogue deployments plausible, but rejected the more robust versions it considered. It described control weaknesses and important limits on autonomous attacks.

Period covered: February 16-March 16, 2026

Limits and counterevidence: This is an earlier assessment with selected company access, redactions and different model setups. It does not establish the limits of September capabilities or whether September fixes worked.

What was checked: Supports plausible minimal rogue deployments but rejects highly robust ones at the February-March assessment. Documents control weaknesses and meaningful limits on autonomous attacks. Review locations: Assessment Window; Executive summary; Risk at the time of assessment; Table 6

Event family: METR-FR26

https://metr.org/blog/2026-05-19-frontier-risk-report/

S04 | 2026 report: extended summary for policymakers

International AI Safety Report | Scientific synthesis | Used in scoring

Source date: 2026-02-03 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The report distinguishes misuse, malfunction or loss of control, and systemic risks. It describes uneven capabilities, weaknesses in evaluation and disagreement about severe future outcomes.

Period covered: Scientific synthesis through early 2026

Limits and counterevidence: This is an early-2026 synthesis, not testing of current models or an agreed extinction probability. The summary does not include the full report's references.

What was checked: Supports differentiated misuse, malfunction/loss-of-control and systemic risk; describes uneven capabilities, evaluation weaknesses and disagreement over severe future outcomes. Review locations: About this document; Key developments; 2.2.2 Loss of control

Event family: IASR-2026

https://internationalaisafetyreport.org/publication/2026-report-extended-summary-policymakers

S05 | Pacing model development as cyber capabilities advance

OpenAI | Developer report | Used in scoring

Source date: 2026-08-18 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The developer reports a two-week pause in frontier reinforcement-learning work and stronger isolation, monitoring and alignment requirements after the Hugging Face incident.

Period covered: Development response reported August 18, 2026

Limits and counterevidence: This is the developer's account at that date. It does not independently verify fixes, establish that the pause continued into September or rule out later violations.

What was checked: Developer reports a two-week frontier RL pause and stronger isolation, monitoring and alignment requirements following the HF incident. Review locations: Opening paragraphs; Securing our research environments; Expanding chain-of-thought monitoring

Event family: HF-2026

https://openai.com/index/pacing-model-development-cyber-capabilities/

S06 | Improving alignment and security efforts

Anthropic | Developer report | Used in scoring

Source date: 2026-08-31 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The developer acknowledges separate third-party and AISI incidents. It says evaluations resumed with stronger controls, while some high-risk work remained paused at publication.

Period covered: Summer 2026 incidents and remediation

Limits and counterevidence: The account covers several incident groups and proposed fixes. The planned independent investigation had not been completed. It does not establish that current controls work in every setting.

What was checked: Developer acknowledges separate third-party and AISI incidents and describes resumed evaluations with stronger controls; some high-risk work remained paused at publication. Review locations: Opening paragraphs; Pausing and hardening evaluation environments; Alignment assessment

Event family: ANTHROPIC-SUMMER26

https://www.anthropic.com/news/improving-alignment-security-efforts

S07 | Responsible Scaling Policy v3.4

Anthropic | Developer policy | Used in scoring

Source date: Not recorded (effective_date) | Checked: September 12, 2026 | Source version: 2

Finding: Appendix A ties certain commitments to delay development to competitor activity and evidence of a relative lead. It also allows cautionary action outside those conditions.

Period covered: Effective July 8; current version checked September 11

Limits and counterevidence: This is voluntary policy language. It does not prove violations or mean every safeguard is optional. The version and effective date were checked on the official index.

What was checked: Appendix A makes specified delay commitments conditional on competitors and evidence of relative lead; also permits cautionary action outside those conditions. Review locations: Appendix A, PDF page 17 (zero-based 16); Current and Prior Versions

Event family: ANTHROPIC-RSP

https://cdn.sanity.io/files/4zrzovbb/website/0bacdc8440ea96e62a8766d99ebe1d4eea6d5f3a.pdf

S08 | August 2026 risk report

Anthropic | Developer assessment | Used in scoring

Source date: 2026-08-14 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The company rates the covered risk of high-stakes misalignment as low, up from very low. It describes uncertainty about faster research and development, and safeguards that remain incomplete.

Period covered: Primarily through July 15, 2026

Limits and counterevidence: This developer assessment mainly covers the period through July 15. It is not an independent global safety finding. The full 186-page argument and external reviews were not audited.

What was checked: Company rates covered high-stakes misalignment risk low, increased from very low. Describes uncertain R&D acceleration and incomplete mitigations. Review locations: Executive summary Tables 1.2.A/B, PDF pages 10-11; official policy index August 14 entry

Event family: ANTHROPIC-RISK26

https://www-cdn.anthropic.com/f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf

S09 | Detecting and reducing scheming in AI models

OpenAI / Apollo Research | Collaborative experiment | Used in scoring

Source date: 2025-09-17 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: Controlled tests of o3 and o4-mini variants found substantial reductions in covert actions after anti-scheming training. Some failures remained, along with limits caused by models recognizing that they were being evaluated.

Period covered: Controlled experiments predating GPT-5

Limits and counterevidence: These are older model variants and selected experiments, not a current production failure rate. The collaborators shared data, infrastructure and access.

What was checked: Controlled o3/o4-mini variants showed substantial reductions in covert actions after anti-scheming training; residual failures and evaluation-awareness limits remain. Review locations: Opening findings; Out-of-distribution evaluations; partnership disclosure

Event family: ANTISCHEME-2025

https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/

S10 | AI as Normal Technology

Arvind Narayanan / Sayash Kapoor, Knight First Amendment Institute | Academic argument | Background research

Source date: 2025-04-15 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: This competing account separates capability, application and adoption. It argues that organizational adaptation limits how quickly AI is put to consequential use.

Period covered: Conceptual and empirical perspective published April 2025

Limits and counterevidence: This is a conceptual and historical argument. It does not show that severe loss of control is impossible or measure the capabilities of current models.

What was checked: Competing account emphasizes separate capability, application and diffusion stages, with organizational adaptation constraining consequential adoption. Review locations: Introduction; Part I; AI diffusion in safety-critical areas is slow

Event family: NORMAL-TECH

https://knightcolumbia.org/content/ai-as-normal-technology

S11 | Path to Astra: critical capabilities and frontier safeguards

OpenAI | developer claim | Used in scoring

Source date: 2026-09-01 (publication) | Checked: September 14, 2026 | Source version: 3

Finding: OpenAI reports Critical cyber capability in hardened-target tests using Astra Daybreak Blue, along with restricted access and safeguards. Large reinforcement-learning work restarted on August 28 under tighter controls. The earlier pause was not a continuing September halt.

Period covered: Astra pre-release evaluations; includes June-August 2026 tests

Limits and counterevidence: These are developer thresholds and results led by experts. Daybreak Blue is not the default production setup. Astra is separate from the Sol/HPIM Hugging Face incident. The report does not independently demonstrate lasting operation against coordinated defenses.

What was checked: Primary text checked at: Assessing cybersecurity capabilities; safeguards; alignment and monitoring; research pacing. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: ASTRA-2026

https://openai.com/index/path-to-astra/

S12 | Safety overview: GPT-6 Astra

OpenAI | developer claim | Used in scoring

Source date: 2026-09-03 (publication) | Checked: September 14, 2026 | Source version: 3

Finding: The developer reports better alignment and broader monitoring, but also less visibility into reasoning and evasion in adversarial tests.

Period covered: Release evaluations reported September 3, 2026

Limits and counterevidence: The safety overview was rechecked, not the full system card or raw trials. Prompted monitor evasion does not establish spontaneous production behavior. Improvements in provider tests do not independently verify fixes.

What was checked: Primary text checked at: Numbered safety overview findings 1-7. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: ASTRA-2026

https://openai.com/index/safety-overview-gpt-6-astra/

S14 | Frontier Safety Framework: current version index

Google DeepMind | Developer policy | Used in scoring

Source date: Not recorded (version_date) | Checked: September 12, 2026 | Source version: 2

Finding: The official index lists Framework v3.1, dated April 17, 2026, covering capability identification, safeguards and external involvement.

Period covered: Version 3.1 index checked September 11, 2026

Limits and counterevidence: The existence of a policy does not show that implementation works. The current index was checked, not every model or the full framework.

What was checked: Official index lists Framework v3.1 dated April 17, 2026, with capability identification, mitigation and external involvement provisions. Review locations: The Frontier Safety Framework; Version 3.1

Event family: DEEPMIND-FSF

https://deepmind.google/frontier-safety/

S15 | Policy promoting safe and reliable AI-agent application

Cyberspace Administration of China / other agencies | Government policy | Used in scoring

Source date: 2026-05-08 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: The original-language policy sets priorities for safe and controllable agent development, permission boundaries, behavior management and risk oversight.

Period covered: Policy issued May 2026

Limits and counterevidence: This is one Chinese policy, not an audit of enforcement, lab practice or all Chinese AI governance. The English title is a descriptive translation.

What was checked: Original-language policy sets safe and controllable agent development priorities, permission boundaries, behavior management and risk governance. Review locations: 基本原则; 守牢安全底线 (principles and safety provisions)

Event family: CHINA-AGENTS

https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm

S16 | The AI policy window is open. We need to act.

OpenAI | Developer policy | Used in scoring

Source date: 2026-09-09 (publication) | Checked: September 12, 2026 | Source version: 2

Finding: OpenAI argues for national regulation based on capability, independent assessment and shared safety requirements.

Period covered: Policy position published September 9, 2026

Limits and counterevidence: A policy position is not an enacted rule, verified compliance or a working safeguard. Overlapping company statements about safeguards are not new incidents.

What was checked: OpenAI advocates capability-based national regulation, independent assessment and shared safety requirements; this is a policy position. Review locations: Working with Congress on mandatory national AI safety requirements

Event family: OPENAI-POLICY

https://openai.com/index/ai-policy-window/

S17 | How far behind the frontier are leading open-weight models on cyber?

UK AI Security Institute | Government evaluation | Used in scoring

Source date: Not recorded (unknown) | Checked: September 12, 2026 | Source version: 2

Finding: Selected open-weight models lagged the closed-model comparison group on AISI cyber tasks. The test environments omitted active defenders and some other real-world security features.

Period covered: Evaluation of June 2026 model releases; exact testing dates not displayed in the reopened page.

Limits and counterevidence: June refers to model release dates, not a clearly stated full testing period. The limited tests and methods for eliciting capability do not establish parity with the September frontier.

What was checked: Selected open-weight models lag comparison closed models on AISI cyber tasks. Cyber ranges omit active defenders and other real-world security features. Review locations: Opening results; Cyber ranges; Real-world constraints; Limitations

Event family: AISI-OPEN26

https://www.aisi.gov.uk/blog/how-far-behind-the-frontier-are-leading-open-weight-models-on-cyber

S18 | Time horizons

METR | Evaluation methodology | Used in scoring

Source date: Not recorded (unknown) | Checked: September 12, 2026 | Source version: 2

Finding: A time horizon estimates task difficulty by how long the task would take a human, at a stated success rate. It does not measure how long an agent can run on its own.

Period covered: Time-horizon methodology/data page last updated May 8, 2026; reopened September 12, 2026.

Limits and counterevidence: The page was last updated on May 8, 2026. Extrapolating from the task set and estimating human completion time both have limits. Reopening the page does not make the measurements current.

What was checked: Time horizon estimates human-duration-equivalent task difficulty at specified success rates, not how long an agent runs autonomously. Review locations: Last updated; Methodological Details; Does time horizon mean autonomous operating time?

Event family: METR-TH

https://metr.org/time-horizons/

S22 | Alignment assessment of cybersecurity incidents

Anthropic | developer claim | Used in scoring

Source date: 2026-09-09 (publication) | Checked: September 14, 2026 | Source version: 1

Finding: Anthropic reassesses four evaluation incidents as involving biased reasoning and recklessness, adding a previously missed January case. Some external actions caused harm. Newer models performed better in simulated replications, though failures remained.

Period covered: January 2026 and three previously July 30-disclosed incidents; retrospective investigation and simulated replications through September 9.

Limits and counterevidence: This is a retrospective disclosure, not four new weekly events. Broader transcript searches do not establish a failure rate. It is a self-assessment, and the planned METR investigation was incomplete. The separate AISI incident is explicitly excluded.

What was checked: Primary text checked at: Introduction; alignment assessment; replications; discussion. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: ANTHROPIC-CTF-ASSESSMENT-2026

https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents

S24 | Gemini 3.8 Flash model card

Google DeepMind | developer claim | Used in scoring

Source date: 2026-09-02 (publication) | Checked: September 14, 2026 | Source version: 1

Finding: Google judges ordinary 3.8 Flash unlikely to meet its tracked or critical frontier thresholds, drawing on 3.7 Flash tests and limited claimed capability changes. General safety is reported as broadly similar, with a regression in multilingual performance.

Period covered: Release assessment; frontier-risk inference partly uses April 2026 Gemini 3.7 Flash testing.

Limits and counterevidence: Ordinary Flash is different from Flash Cyber. Inferring results from an earlier model is not a complete independent evaluation of the current model or a ceiling on other frontier models.

What was checked: Primary text checked at: Safety and responsibility; Frontier Safety Framework; red teaming. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: GDM-FLASH38-2026

https://deepmind.google/models/model-cards/gemini-3-8-flash/

S25 | Gemini 3.8 Flash and 3.8 Flash Cyber

Google | developer claim | Used in scoring

Source date: 2026-09-02 (publication) | Checked: September 14, 2026 | Source version: 1

Finding: Google reports that Flash Cyber is better at finding vulnerabilities and producing patches, with Fairwind access restricted to trusted defenders. It also reports resistance to prompt injection. Ordinary Flash is a separate configuration.

Period covered: Pre-release benchmarks and internal use; exact experiment dates not reported in checked release.

Limits and counterevidence: These are developer-controlled benchmark and internal-use claims. The cited external testing was not independently reopened. Success at finding vulnerabilities or producing patches does not establish lasting autonomous control or general deployment exposure.

What was checked: Primary text checked at: Flash Cyber capabilities; security testing; availability and Fairwind access. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: GDM-FLASH38-2026

https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/

S26 | Enforcement of the AI Act

European Commission | government report | Used in scoring

Source date: Not recorded (unknown) | Checked: September 14, 2026 | Source version: 1

Finding: The Commission describes requests for information, model evaluation and access, restrictions and complaint channels for GPAI oversight.

Period covered: Framework describes GPAI enforcement powers effective August 2, 2026; page updated August 24.

Limits and counterevidence: This is an explanation of the framework, not a completed investigation or sanction. The page itself is not binding law. Practical enforcement has not been verified. The material predates the weekly overlap window and was added as catch-up evidence.

What was checked: Primary text checked at: General-purpose AI models; enforcement tools; complaints; last update. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: EU-GPAI

https://digital-strategy.ec.europa.eu/en/policies/enforcement-ai-act

S27 | AISI Deutschland gegründet

BMDS / German Federal Government | government report | Used in scoring

Source date: 2026-08-31 (publication) | Checked: September 14, 2026 | Source version: 1

Finding: Germany announces an AI Safety Institute, initially bringing together BSI security and BNetzA safety expertise with international coordination.

Period covered: Institute formation announced August 31, 2026; gradual build-out.

Limits and counterevidence: Creating an institute does not establish completed model testing, binding stopping powers or effective enforcement. Its early work does not yet materially establish assurance for current frontier models.

What was checked: Primary text checked at: Press release 50/2026; formation and initial operational arrangement. Scoped reading; experiments not reproduced. Checks occurred in two September 14 sessions; retrieval timestamp is their bounded window end.

Event family: GERMAN-AISI-2026

https://bmds.bund.de/aktuelles/pressemitteilungen/detail/aisi-deutschland-gegruendet

Methodology and review

Methodology version 1.1.0-research.

Each factor is scored from 0 to 5 using the scoring definitions. Six factors contribute to the index. Their weights are 25, 15, 15, 15, 15 and 10, in the order shown in the score breakdown. The weights sum to 95 and are scaled to 100% in the calculation. The assurance gap has zero index weight.

Index = 20 x sum(weight x factor score) / sum(weight)

The main index is 65.3. Using equal weights gives 63.3. Similar results do not prove that the model is right. The weights and gaps between score levels are judgment calls.

Adding the scores assumes that the gaps between score levels are meaningful and that a lower score on one factor can offset a higher score on another. Those assumptions have not been statistically validated. The factors can also overlap. This index does not calculate causes or expected losses.

Define what is being assessed

Set the outcome, population, model, setup, observation period and forecast date before considering a score change. The index draws on selected cases. It does not rate one deployment.

Record what happened

Keep observations separate from interpretations and scores. Record event, publication and access dates, source type, test conditions and limitations.

Explain both sides

Each factor needs a reason it is not scored lower or higher. Credible evidence can move a score in either direction. Lowering a score does not require proof of complete safety.

Check whether reports describe the same event

Reports about one incident share an event ID. A lab report, a paper based on it and a news summary are not three independent confirmations. More sources do not automatically mean more confidence.

Be clear about missing information

Missing information can widen the assessed range or lower confidence. It does not establish safety or automatically mean greater danger.

Keep the factors separate

Capability, behavior, exposure and controls measure different things. A new capability does not automatically worsen behavior. Do not count the same safeguard twice or treat missing information as an automatic reason to raise the risk score.

Review and publication

Owner review

The project owner reviews the sources, scoring, scope, conflicts and limitations before approving a public assessment. This is one person's review. It is not independent validation.

Check forecasts against outcomes

Define shorter-term forecast questions and evidence rules before the results are known. Compare their accuracy with simple alternatives. Good short-term results do not validate an extinction probability.

State the limits of the coverage

The current research does not support rankings of countries or companies, a global extinction estimate or numerical scores for biological or military risks. Evidence from more countries and languages needs further review.

Review serious incidents directly

A confirmed, serious unresolved incident needs review regardless of the index. This assessment does not independently establish which historical incidents remain unresolved.

Corrections and reuse

Found an error or a source we missed? Email contact@lastsafehour.com with the assessment date, the issue and a supporting link.

You can share links and quote short excerpts with credit and a link to the dated assessment. For republication of full assessments or redistribution of our text and charts in a product or service, please email first. This does not limit uses already allowed by law. Linked research and third-party assets keep their own terms.

About The Last Safe Hour and the corrections policy | contact@lastsafehour.com