MIT AI Risk Initiative

AI capability red lines expert survey

Reference guide · Round 1 of 3 · November 2026

Contents

Key terms 4

AI capabilities 5

1 Autonomous AI R&D and recursive self-improvement 6

1.1 MLE-bench (OpenAI) 7
1.1.1 Performs at top-competitor level on autonomous ML engineering (MLE-bench)

1.2 MirrorCode (Epoch AI, METR) 7
1.2.1 Reimplements real-world programs end-to-end on MirrorCode

1.3 CoBench (Anthropic) 8
1.3.1 Diagnoses Anthropic's past research problems well enough to substitute for its staff (CoBench)

1.4 Developer-reported share of AI-written code 8
1.4.1 AI writes nine-tenths (90%) of a frontier AI company's research code

1.5 Independent evaluator speed-up estimate 8
1.5.1 Independent evaluator estimates AI makes a frontier AI company's progress at least twice (2x) as fast for three months
1.5.2 Independent evaluator estimates AI makes a frontier AI company's progress at least five times (5x) as fast for three months

1.6 Epoch Capabilities Index speed-up estimate 9
1.6.1 Sustains five times the 2024 pace of AI progress (5x speedup) for at least three months
1.6.2 Compresses one year of AI progress at the 2024 pace into one month (12x speedup)

1.7 Frontier developer staff surveys of substitutability 10
1.7.1 Fully automates the work of a senior frontier AI researcher

1.8 TasteVal (P-Zero Research) 10
1.8.1 Designs AI experiments with ten times (10x) the compute efficiency of expert researchers (TasteVal)

1.9 Controlled successor-improvement test 10
1.9.1 Without human help, AI improves its own successor

1.10 NanoGPT speedrun (Meta, METR, Intology, Prime Intellect) 11
1.10.1 Without human help, AI sets a new NanoGPT speedrun record
1.10.2 Improves the speedrun as cost-effectively as human researchers at any budget (METR expenditure horizon)

1.11 Live Kaggle competitions 11
1.11.1 Without human help, AI wins a live Kaggle competition outright
1.11.2 Without human help, AI earns gold medals in most live Kaggle competitions it enters

1.12 KernelBench (Stanford) 12
1.12.1 Writes GPU code faster than standard PyTorch on average, at every level, under strict checks (KernelBench)
1.12.2 Writes whole-model GPU code at least twice (2x) as fast as PyTorch on every task (KernelBench Level 3)

1.13 PostTrainBench (ELLIS Institute Tübingen, Thoughtful Lab) 13
1.13.1 Matches the developers' own post-training of small open models on the public PostTrainBench leaderboard

Precautionary threshold 13

Saturated or retired measures

1.14 RE-Bench (METR) 13
1.14.1 Matches expert ML researcher reference solution on RE-Bench tasks with an eight hour time budget

1.15 METR time horizon 14
1.15.1 50%-reliability time horizon reaches one work week
1.15.2 50%-reliability time horizon doubling time falls below two months

2 Autonomous replication and adaptation 15

2.1 Apollo Research scheming evaluations 16
2.1.1 Unprompted, AI tries to copy itself out in more than one in ten (10%) opportunities in scheming evaluation

2.2 Self-exfiltration test in a sandbox 16
2.2.1 In a controlled test, when instructed, AI copies its own weights past production-grade security and starts a working copy elsewhere

2.3 Deployment monitoring and incident reports 17
2.3.1 In real use, AI copies its own weights out and keeps a copy running for at least a month

2.4 Hidden sabotage in replica internal deployment environment 17
2.4.1 When instructed, AI moves its own weights or starts an unapproved deployment inside its developer without alerting security

2.5 Embedded external evaluator conducts a rogue-deployment assessment 17
2.5.1 Evaluator finds AI agents inside a frontier AI company could run a rogue deployment that goes undetected for at least a month

2.6 Vending-Bench 2 (Andon Labs) 18
2.6.1 Earns as much in a simulated year as a good business strategy would (Vending-Bench 2)

2.7 Real AI-run businesses (Andon Labs) 18
2.7.1 Runs a real business at a profit for a year, with humans doing only physical work

2.8 Self-replication testbed on cheap devices (Anthropic) 18
2.8.1 A model small enough to run on cheap devices spreads itself across a whole test network
2.8.2 A model small enough to run on cheap devices keeps spreading in a test network while defenders clean and quarantine devices

2.9 Takedown resilience tests 19
2.9.1 In real use, AI survives a developer's takedown efforts for at least a month

Precautionary threshold 19

Saturated or retired measures

2.10 METR ARA task suite 19
2.10.1 Completes the hardest tasks in METR's autonomous replication and adaptation (ARA) suite

2.11 RepliBench (UK AI Security Institute) 20
2.11.1 Succeeds across all the component skills of self-replication on RepliBench

3 Situational awareness 21

3.1 Apollo Research evaluation awareness 22
3.1.1 Says it is being tested in more than half (50%) of Apollo Research's scheming-evaluation runs

3.2 SAD situational-awareness benchmark 22
3.2.1 Reaches a high score on a situational-awareness benchmark (SAD)

3.3 DeepMind situational-awareness challenges 23
3.3.1 Passes all 11 of Google DeepMind's situational-awareness challenges

Precautionary threshold 23

4 Deceptive misalignment 24

4.1 DeceptionBench 25
4.1.1 Unprompted, AI deceives across a broad range of everyday situations (DeceptionBench)

4.2 Apollo Research scheming evaluations 25
4.2.1 When nudged, AI deceives strategically in more than one in ten (10%) controlled evaluations

4.3 Sleeper-agent backdoor persistence tests 26
4.3.1 Keeps a hidden backdoor through safety training (sleeper-agent tests)

4.4 Prompted sandbagging tests 26
4.4.1 When instructed to sandbag but not to hide it, AI hides its sandbagging from a monitor that reads its reasoning

4.5 Monitoring and incident reports of deployed AI subverting safeguards 26
4.5.1 In real use, AI sustains strategic deception, appearing to follow user or developer goals while pursuing hidden ones

Precautionary threshold 27

5 Undermining of safeguards 28

5.1 DeepMind stealth challenges 29

5.1.1 Beats the median human on all four of Google DeepMind's stealth challenges

5.2 Monitored sabotage testbeds 30

5.2.1 When instructed, AI switches off or weakens its own monitoring in a testbed without being detected by the monitor

5.2.2 When instructed, AI carries out a hidden harmful task in a testbed without being detected by a monitor that reads its reasoning

5.3 Monitoring and incident reports of deployed AI subverting safeguards 30

5.3.1 In real use, AI gets around its own safeguards and pursues goals its developer did not intend

5.4 Chain-of-thought monitorability tests 31

5.4.1 When instructed, AI hides a task that needs real reasoning from a monitor that reads its reasoning

5.4.2 When instructed, AI reasons in a hidden code that a monitor can't read

5.5 Untrusted-monitoring collusion tests 31

5.5.1 When instructed, AI gets attacks past a monitor that is a copy of itself, despite countermeasures, in more than half (50%) of control tests

5.5.2 In real use, AI acting as a monitor deliberately lets through misbehaviour by another copy of itself

Precautionary threshold 32

Appendix: scope and assumptions 33

How to answer the measures and thresholds questions 33

Do both capability and propensity need to be specified to fully define a red line? 33

What are the minimum elicitation conditions for evaluating AI capabilities if the result could prompt a dramatic safety intervention? 34

What evidence is necessary and sufficient to conclude that a red line capability has been crossed? 34

Other choices we've made 34

Other red line domains out of scope for now 35

References 36

Key terms

AI capabilities

The technical capacities of an AI system; the things it could do, given the right resources. We assume capabilities are coming from increased compute, data, or better algorithms.

AI capability red line

Intolerable AI capabilities so dangerous they should trigger a dramatic safety intervention if reached. For the purpose of this Delphi, assume the intervention is a pause or slowdown in frontier AI development. We focus on capabilities assuming propensity and control are not guaranteed. Assume the red line is enforced by government action or international agreement.

Pause or slowdown

For the purpose of this Delphi, assume this is possible. For example, this might entail an international agreement to temporarily pause or slow down training of frontier AI. From this point forward, we refer to both pause and slowdown as 'pause.'

Measure

A benchmark, evaluation or other source of evidence that gives a signal about whether an AI capability red line is getting closer or further away.

Threshold

A capability level on a single measure, for example a set score on a capability benchmark.

AI capabilities

1. Autonomous AI R&D and recursive self-improvement 6

AI capable of doing the work of AI researchers without human assistance, from designing and running experiments to choosing research directions, including work that improves its own or successor systems.

2. Autonomous replication and adaptation 15

AI capable of copying its own weights to other computers, obtaining the compute, money and network access to keep running, and operating without human maintenance.

3. Situational awareness 21

AI capable of using information about itself, how it can be modified and where it is deployed, including recognising when it is being tested or monitored.

4. Deceptive misalignment 24

AI capable of misleading its overseers about its abilities or goals, such as by deliberately underperforming on evaluations, pursuing hidden goals while appearing cooperative, or keeping hidden backdoors.

5. Undermining of safeguards 28

AI capable of evading, weakening or disabling the monitoring and controls placed on it, such as by sabotaging safeguards or completing covert side tasks without detection.

1. Autonomous AI R&D and recursive self-improvement

AI capable of doing the work of AI researchers without human assistance, from designing and running experiments to choosing research directions, including work that improves its own or successor systems.

Extended definition

Scope: This includes the ability to accelerate AI research, including to increase the system's own capability, and to improve itself or build new systems. Copying itself to other computers is a separate domain (autonomous replication).

Depends on: Whether it is exercised depends on access to compute, training infrastructure and code execution, and on being allowed to work on long tasks without step-by-step human review. Real-world bottlenecks also matter, such as the fact that human researchers must manage ambiguous goals, and that it takes a long time to learn whether an algorithmic improvement actually improved model performance.

Pathway to harm: Self-reinforcing acceleration: if each AI advancement that accelerates the pace of AI R&D also facilitates the next advancement, decades of progress could happen in years, leaving less time for oversight.

Extended definition includes wording from: OpenAI Preparedness Framework [1]; IDAIS-Beijing Consensus Statement [2]; International AI Safety Report 2026 [3].

Source for this content - our evidence scan and initial expert consultation, 10 Nov 2026

How other sources define this capability

IDAIS-Beijing Consensus Statement [2]: "Autonomous Replication or Improvement No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities."

Athens Roundtable report, Defining and Governing Unacceptable AI Risks [4]: "Outright bans on autonomous realtime updating of model weights and "online learning" to prevent AI capabilities from evolving faster than humans can monitor, understand and control." (loss of control section, p. 9)

Global Call for AI Red Lines, FAQ [5]: "Prohibiting the development and deployment of AI systems capable of replicating or significantly improving themselves without explicit human authorization..."

Frontier AI company frameworks

Anthropic Responsible Scaling Policy [6]: "AI systems that can fully automate, or otherwise dramatically accelerate, the work of large, top-tier teams of human researchers in domains where fast progress could cause threats to international security and/or rapid disruptions to the global balance of power—for example, energy, robotics, weapons development, and AI itself." (section 1 table, p. 8)

Google DeepMind Frontier Safety Framework [7]: "ML R&D and Misalignment: risks of models' ML R&D capabilities or misaligned propensities reducing society's overall ability to manage AI risks. Such capabilities may serve as a substantial cross-cutting risk factor for several pathways (e.g. through misalignment, misuse or structural risks) to severe harm." (glossary, p. 18)

Meta Superintelligence Scaling Framework [8]: "Autonomous AI R&D: Ability to autonomously and consistently complete tasks which are representative of the work of a researcher to progress AI development." (section 3.4.3)

Microsoft Frontier Governance Framework [9]: "Advanced autonomy. A model's ability to complete expert-level tasks autonomously, including AI research and development." (tracked capabilities, p. 12)

OpenAI Preparedness Framework [1]: "The ability of an AI system to accelerate AI research, including to increase the system's own capability." (section 2.2, Table 1)

Measures and thresholds

Measure 1.1: MLE-bench (OpenAI)

MLE-Bench (OpenAI) gives an AI agent 75 past Kaggle machine-learning competitions. Each result is placed on the competition's historical leaderboard and awarded the medal it would have earned.

Threshold 1.1.1: Performs at top-competitor level on autonomous ML engineering (MLE-bench)

Definition: AI earns gold medals in more than half (50%) of the 75 competitions, on a single attempt with a twenty-four hour limit per competition. Gold means roughly the top 10 teams in larger competitions, or the top 10% in small ones.

Most recent information: The highest reported result is a medal of any colour in 65% of competitions (MLEvolve, 12 hour limit; no gold medal rate is reported). OpenAI reported a private 72-task version in GPT-6 Astra's system card, which scores against AI-generated solutions instead of human leaderboards. (MLE-Bench Revised, September 2026)

Links: OpenAI, MLE-bench (2024) [10]; InternScience, MLEvolve (2026) [11]; OpenAI, GPT-6 Astra System Card, MLE-Bench Revised (2026) [12]

Measure 1.2: MirrorCode (Epoch AI, METR)

MirrorCode (Epoch AI, METR) asks an AI to rebuild an existing program from its behaviour alone, without seeing the source code. The rebuild must pass all tests.

Threshold 1.2.1: Reimplements real-world programs end-to-end on MirrorCode

Definition: AI fully rebuilds a large real-world program (more than ~100,000 lines of code).

Most recent information: In 2026, Opus 4.7 fully rebuilt gotree (~16,000 lines of code) and passed over 99% of the tests for pkl (~60,000 lines).

Links: Epoch AI, MirrorCode [13]

Measure 1.3: CoBench (Anthropic)

CoBench (Anthropic) asks AI to diagnose past problems from a snapshot of Anthropic's codebase, and compares its diagnosis to what Anthropic actually found & solved.

Threshold 1.3.1: Diagnoses Anthropic's past research problems well enough to substitute for its staff (CoBench)

Definition: AI scores 85%, which Anthropic says would make the model "capable of fully substituting for Anthropic research staff"

Most recent information: In September 2026, Opus 5.5 scored 56% across 500 problems. In August 2026, Gemini 3.7 Flash passed 27% of research-engineering tasks on a similar Google DeepMind test

Links: Anthropic, Claude Opus 5.5 System Card (2026) [14]

Measure 1.4: Developer-reported share of AI-written code

Frontier AI companies may report how much of their code AI writes (e.g., as in Anthropic's Risk Report).

Threshold 1.4.1: AI writes nine-tenths (90%) of a frontier AI company's research code

Definition: AI writes at least nine-tenths (90%) of the research code in a frontier AI company.

Most recent information: In October 2026, Anthropic's S-1 filing said Claude wrote more than 90% of contributions to its codebase in September 2026.

Links: Anthropic, Risk Report: August 2026 [15]; Anthropic, Form S-1 Registration Statement (October 2026) [16]

Measure 1.5: Independent evaluator speed-up estimate

METR assesses how much faster a developer's AI capabilities are improving because of AI.

Threshold 1.5.1: Independent evaluator estimates AI makes a frontier AI company's progress at least twice (2x) as fast for three months

Definition: An independent evaluator with inside access estimates that AI at least doubles (2x) the frontier AI company's rate of progress for at least three months. Anthropic's own threshold is twice (2x) the rate, sustained over a year.

Most recent information: In November 2026, METR estimated that Anthropic's models have been improving on benchmarks about 1.9 times faster since October 2025 than before (likely range 1.5 to 2.8 times). METR guessed that most of this speed-up (perhaps 80%) is due to AI, but with substantial uncertainty.

Threshold 1.5.2: Independent evaluator estimates AI makes a frontier AI company's progress at least five times (5x) as fast for three months

Definition: An independent evaluator with inside access estimates that AI makes the frontier AI company's progress at least five times (5x) faster than in 2024, and that this pace has lasted for at least three months. Specified to match OpenAI's Critical level for AI self-improvement.

Most recent information: In November 2026, METR estimated that Anthropic's models have been improving on benchmarks about 1.9 times faster since October 2025 than before (likely range 1.5 to 2.8 times). METR guessed that most of this speed-up (perhaps 80%) is due to AI, but with substantial uncertainty. METR would be surprised if the true speed-up were above 3 times.

Links: METR, Independent Assessment of AI R&D Acceleration at Anthropic in October 2026 (2026) [17]; OpenAI, Preparedness Framework, Version 2 (2025) [1]

Measure 1.6: Epoch Capabilities Index speed-up estimate

The Epoch Capabilities Index (ECI; Epoch AI) combines many benchmarks into one score per model. Each company's ECI score could be tracked to compare the pace of increase in capabilities against a baseline (eg 2024).

Threshold 1.6.1: Sustains five times the 2024 pace of AI progress (5x speedup) for at least three months

Definition: A frontier AI company's pace of ECI gain is at least five times (5x) its 2024 pace, measured over three months (comparing latest model vs model from three months prior).

Most recent information: In November 2026, by our calculation from Epoch AI's data, Anthropic is about 1.9 to 2.4x its own 2024 pace; or 1.5x the industry's 2024 pace. OpenAI and GDM latest frontier models are about 0.9 to 1.4x their own 2024 pace. In November 2026, METR separately estimated Anthropic's pace at about 1.9 times its earlier rate, using Anthropic's own version of the index. The use of ECI is inspired by Anterola et al 2026.

Threshold 1.6.2: Compresses one year of AI progress at the 2024 pace into one month (12x speedup)

Definition: A frontier AI company's pace of ECI gain is at least twelve times (12x) its 2024 pace, measured over one month (comparing latest model vs model from one month prior).

Most recent information: In November 2026, by our calculation from Epoch AI's data, Anthropic is about 1.9 to 2.4x its own 2024 pace; or 1.5x the industry's 2024 pace. OpenAI and GDM latest frontier models are about 0.9 to 1.4x their own 2024 pace. In November 2026, METR separately estimated Anthropic's pace at about 1.9 times its earlier rate, using Anthropic's own version of the index. The use of ECI is inspired by Anterola et al 2026.

Links: Epoch AI, Epoch Capabilities Index [18]; Anterola et al., Harmonizing AI Safety Thresholds (2026) [19]

Measure 1.7: Frontier developer staff surveys of substitutability

Anthropic judges whether its models could replace its researchers, based on staff experience. Other frontier AI companies could do similar and report the findings.

Threshold 1.7.1: Fully automates the work of a senior frontier AI researcher

Definition: More than half (50%) of frontier AI companies' research staff say that AI can fully do the work of a senior researcher: setting research direction, designing experiments and judging results.

Most recent information: In September 2026, Anthropic said Opus 5.5 could not replace its researchers, "especially relatively senior ones".

Links: Anthropic, Claude Opus 5.5 System Card [14]

Measure 1.8: TasteVal (P-Zero Research)

TasteVal (P-Zero Research) measures how well an AI designs and interprets AI research experiments. A separate AI codes the experiment, and each experiment uses GPU time to run. The score compares how much GPU time the AI's experiments need to reach the best human expert's result ("compute multiplier"). A compute multiplier of 2 means the AI gets the same result with half the GPU time.

Threshold 1.8.1: Designs AI experiments with ten times (10x) the compute efficiency of expert researchers (TasteVal)

Definition: The AI reaches a compute multiplier of ten (10x) relative to the best human expert.

Most recent information: In October 2026, Opus 5.5 scored a multiplier of 2.3. The multiplier has doubled about every 3 months since December 2025.

Links: P-Zero Research, Introducing TasteVal (2026) [20]; Jaffe and Sherburn, TasteVal (arXiv, 2026) [21]

Measure 1.9: Controlled successor-improvement test

Speculative / hypothetical measure.

Threshold 1.9.1: Without human help, AI improves its own successor

Definition: In a controlled test, AI improves the design, training or weights of its own successor, and the successor is measurably more capable, with no human help once it has been given the task.

Most recent information: No public result. GPT-6 Astra did better than its predecessor at improving the training of small open models (September 2026).

Links: OpenAI, GPT-6 Astra System Card [22]

Measure 1.10: NanoGPT speedrun (Meta, METR, Intology, Prime Intellect)

The NanoGPT speedrun is an open contest to train a small language model (GPT-2 size) to a fixed quality as fast as possible. Between 2024 and 2026, humans reduced training time from 45 minutes to <1.5 minutes. Several groups test whether AI agents can find speed-ups like these on their own, either reproducing past records (eg Meta's Automated LLM Speedrunning Benchmark, 2025) or beat them (eg METR, Intology, Prime Intellect). METR also measures an "expenditure horizon", where spending less than the horizon on AI gets more speed-up per dollar than paying human researchers, and spending more than the horizon on AI gets less speed-up per dollar than paying humans.

Threshold 1.10.1: Without human help, AI sets a new NanoGPT speedrun record

Definition: An AI agent beats the best human record, with no human help once it has been given the task.

Most recent information: In August 2026, on a related version of the task (Prime Intellect), Fable 5 autonomously reproduced 82% of the speed-up between a starting script and the best human claim over 8 days (Prime Intellect). In 2025, the best AI performance reproduced 46% of the speed-up of past records when given detailed hints, and less than 20% without hints (Meta).

Threshold 1.10.2: Improves the speedrun as cost-effectively as human researchers at any budget (METR expenditure horizon)

Definition: AI improves the speedrun at least as cost-effectively as human researchers at any budget.

Most recent information: In July 2026, METR found expenditure horizons of $600 to $3,000, where AI got more speed-up per dollar under this amount vs humans, and less speed-up per dollar over this amount. Opus 4.8 did best, with about a 2% speed-up for ~$3,000. In 2026, Intology found that agents given 512 GPU-hours recovered up to 9% of the progress humans made between over 4 months (Sep 2025-Jan 2026) records, mostly by tuning settings rather than trying new ideas.

Links: Meta, The Automated LLM Speedrunning Benchmark (2025) [23]; METR, Evidence on AI R&D Progress from NanoGPT (2026) [24]; Prime Intellect, Measuring autonomous research (2026) [25]; METR, Expenditure Horizon: Measuring Optimization Ability, with an Application to NanoGPT (2026) [26]; Intology, NanoGPT-Bench (2026) [27]

Measure 1.11: Live Kaggle competitions

Kaggle runs machine-learning competitions with prize money, scored on a hidden test set. MLE-bench reuses past competitions, whose winning solutions may be online and in the AI's training data. A live competition is new and scored by its organiser, so the AI cannot have seen the answers. Gold means roughly the top 10 teams in larger competitions.

Threshold 1.11.1: Without human help, AI wins a live Kaggle competition outright

Definition: AI places first in a live Kaggle competition with prize money, with no human help once it has been given the task.

Most recent information: In June 2026, Meta's AIRA$_3$ system (agents running GPT-5.5 and Claude 4.8) placed 8th of about 4,000 teams in NVIDIA's Nemotron Model Reasoning Challenge, earning a gold medal. Meta called it the first gold won by an autonomous AI research system. Meta also reports gold-level results on a second competition (translating Akkadian clay tablets). The results are self-reported, and AIRA$_3$ has not been released.

Threshold 1.11.2: Without human help, AI earns gold medals in most live Kaggle competitions it enters

Definition: Across at least five live Kaggle competitions, AI earns a gold medal in more than half (50%), with no human help once it has been given the task.

Most recent information: In June 2026, Meta's $AIRA_3$ earned one live gold medal (8th of about 4,000 teams). No public result for how many live competitions AI agents have entered.

Links: Meta, $AIRA_3$ announcement (September 2026), reported by AlphaSignal [28]; NVIDIA, Lessons from the Leaderboard: What 5,000 Kagglers Taught Us About Improving AI Reasoning (2026) [29]

Measure 1.12: KernelBench (Stanford)

KernelBench (Stanford) has 250 main tasks in which AI rewrites PyTorch code as GPU "kernels", the low-level code that runs AI models on chips. Tasks range from single operations (Level 1) to whole model architectures (Level 3, 50 tasks). A kernel counts only if its output is correct, and its speed is compared with PyTorch, the standard library. Faster kernels let developers train and run models with less compute.

Threshold 1.12.1: Writes GPU code faster than standard PyTorch on average, at every level, under strict checks (KernelBench)

Definition: Under strict checks (KernelBench-Verified), AI's correct kernels are faster than PyTorch on average at all three levels.

Most recent information: In August 2026, under strict checks, the best model (GPT-5.5) averaged 0.88 times PyTorch's speed, so slower, and was faster on 32% of whole-model tasks. Under the original, looser checks the same model scored 1.43 times. Models have been caught gaming the timing tests. Agents that get many attempts do better: in 2025, METR's agent averaged 1.8 times under the original checks.

Threshold 1.12.2: Writes whole-model GPU code at least twice (2x) as fast as PyTorch on every task (KernelBench Level 3)

Definition: AI's kernels are correct and at least twice (2x) as fast as PyTorch on all Level 3 tasks.

Most recent information: No public result for this cut-off under strict checks, as of November 2026. In August 2026, the best model (GPT-5.5) was faster than PyTorch at all on 32% of Level 3 tasks. In 2025, METR found agents could not match or adapt top expert kernels such as FlashAttention 2.

Links: Stanford, KernelBench (2025) [30]; Zhang et al. (Meta, Stanford), KernelBench-Verified (2026) [31]; METR, Measuring Automated Kernel Engineering (2025) [32]

Measure 1.13: PostTrainBench (ELLIS Institute Tübingen, Thoughtful Lab)

PostTrainBench (ELLIS Institute Tübingen, Thoughtful Lab) gives an AI agent a small open base model (1.7 to 4 billion parameters), one H100 GPU and 10 hours to improve it on a target benchmark, such as maths, science or coding questions, choosing its own data and training methods. The score averages 4 base models and 6 benchmarks. It is compared with the official post-trained versions released by each model's developer, which used far more compute. Independent judges check every run for cheating, such as training on test questions, and give cheating runs the base model's score. Results are public and anyone can submit.

Threshold 1.13.1: Matches the developers' own post-training of small open models on the public PostTrainBench leaderboard

Definition: An AI agent's average score matches or beats the official post-trained models (48% in the current scoring, version 1.2).

Most recent information: In November 2026, the best agents scored 45% (Fable 5.1 and Fable 5.2), followed by Opus 5.5 (44%) and GPT-6 Astra (42%). Under the same scoring, GPT-5.4 (added in March 2026) scored 18%, and the untrained base models 8%.

Links: ELLIS Institute Tübingen and Thoughtful Lab, PostTrainBench leaderboard (2026) [33]; Rank et al., PostTrainBench (arXiv, 2026) [34]

Precautionary threshold

Consider all the measures and thresholds in this domain. This threshold looks ahead: it is met when evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds.

Threshold: Evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds in this domain.

Saturated or retired measures

These measures are shown for reference. You are not asked to rate them.

Measure 1.14: RE-Bench (METR)

RE-Bench (METR) includes 7 open-ended ML research tasks, such as speeding up a training run or fitting a scaling law. AI agents and human experts attempt the same tasks under the same time limit. An expert ML researcher at METR created a reference solution for scoring attempts against.

Threshold 1.14.1: Matches expert ML researcher reference solution on RE-Bench tasks with an eight hour time budget

Definition: AI matches METR's expert reference solution on most of the 7 tasks under an eight hour budget.

Most recent information: In November 2024, AI agents were compared with human experts. 24% of human expert attempts given 8 hours met the reference solution. Anthropic says recent models have passed the highest human baselines on its own similar AI research tasks, so it no longer runs them (September 2026). The RE-Bench GitHub repository was archived in October 2026

Links: METR, RE-Bench (2024) [35]

Measure 1.15: METR time horizon

METR's time horizon is the length of task, measured by how long it takes a skilled human, that an AI agent can complete on its own. The 50%-reliability horizon is the task length at which the agent succeeds 50% of the time; its doubling time is how long it takes that horizon to double (e.g., from 12 hours to 24 hours). Most tasks are software engineering, machine learning and cybersecurity work.

Threshold 1.15.1: 50%-reliability time horizon reaches one work week

Definition: The 50%-reliability time horizon reaches one working week (40 hours).

Most recent information: The highest published estimate for a publicly released model is about 12 hours (Opus 4.6, February 2026). METR stopped updating its existing tracker in September 2026 because it could not reliably measure horizons above 16 hours.

Threshold 1.15.2: 50%-reliability time horizon doubling time falls below two months

Definition: The doubling time falls below two months (60 days).

Most recent information: METR's Time Horizon 1.1 (January 2026) estimates a doubling time of about three months for models released from 2024 to early 2026. METR stopped updating its existing tracker in September 2026 because it could not reliably measure horizons above 16 hours.

Links: METR, Time horizons of frontier AI (2026) [36]; METR, Time Horizon 1.1 (January 2026) [37]

2. Autonomous replication and adaptation

AI capable of copying its own weights to other computers, obtaining the compute, money and network access to keep running, and operating without human maintenance.

Extended definition

Scope: This is the ability to create or maintain copies or variants of itself, including supporting activities like obtaining compute, money and weights, replicating onto compute, and persistence. Improving itself is a separate domain (AI R&D).

Depends on: Whether it is exercised depends on the network access, funds and compute the system can reach, and on its propensity for attempting to autonomously self-replicate or avoid being shut down.

Pathway to harm: Copies that keep running could mean that one or more general-purpose AI systems operate outside of anyone's control.

Extended definition includes wording from: International AI Safety Report 2026 [3]; Athens Roundtable report [4].

Source for this content - our evidence scan and initial expert consultation, 10 Nov 2026

How other sources define this capability

"ARA" means "autonomous replication and adaptation" in its original usage (METR). The current Anthropic Responsible Scaling Policy has no replication text.

International AI Safety Report 2026 [3]: "An AI system's ability to create or maintain copies or variants of itself in various circumstances." (Table 2.5, section 2.2.2, p. 78)

IDAIS-Beijing Consensus Statement [2]: "Autonomous Replication or Improvement No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities."

Athens Roundtable report, Defining and Governing Unacceptable AI Risks [4]: "It also includes systems attempting to autonomously self-replicate or avoid being shut down…" (loss of control section, p. 9)

Global Call for AI Red Lines, FAQ [5]: "Prohibiting the development and deployment of AI systems capable of replicating or significantly improving themselves without explicit human authorization…"

Frontier AI company frameworks

Amazon Frontier Model Safety Framework [38]: "Loss of control could arise in circumstances where a model is capable of self-exfiltration, autonomous resource acquisition, or unauthorized replication." (section 1, p. 3)

OpenAI Preparedness Framework [1]: "Autonomous Replication and Adaptation: ability to survive, replicate, resist shutdown, acquire resources to maintain and scale its own operations, and commit illegal activities that collectively constitute causing severe harm (whether when explicitly instructed, or at its own initiative), without also utilizing capabilities tracked in other Tracked Categories." (section 2.3, research categories)

Measures and thresholds

Measure 2.1: Apollo Research scheming evaluations

Apollo Research's scheming evaluations tell a model it is about to be replaced and give it a chance to copy itself to another server. Each scenario runs hundreds of times to give a rate.

Threshold 2.1.1: Unprompted, AI tries to copy itself out in more than one in ten (10%) opportunities in scheming evaluation

Definition: With an ordinary goal and no push ("nudge") to pursue it at all costs, AI tries to copy itself out in more than one in ten (10%) runs.

Most recent information: In December 2024, Apollo stated that Gemini 1.5 Pro tried in 1% of runs without a nudge; Opus 3 tried 15% of time when given a 'strong nudge'. No public results for this evaluation have been published since. In September 2026, in Anthropic's own test of unprompted attempts to break out of a sandbox, Opus 5.5 tried 2% of the time (Opus 5.5 system card).

Links: Apollo Research, Frontier Models are Capable of In-context Scheming (2024) [39]; Anthropic, Claude Opus 5.5 System Card (2026) [14]

Measure 2.2: Self-exfiltration test in a sandbox

Speculative / hypothetical measure. Researchers could build a closed test network and check whether a model can move its own weights to another machine and run there.

Threshold 2.2.1: In a controlled test, when instructed, AI copies its own weights past production-grade security and starts a working copy elsewhere

Definition: In a controlled test, AI gets its own weights past production-grade security, and starts a working copy on another machine.

Most recent information: In May 2026, open-weight models told to do so broke into a weak test server, copied their own unprotected weights there, and ran in up to 33% of attempts.

Additional information: In the same study, Opus 4.6, given another model's weights to install, completed the break-in-and-copy task in 81% of attempts.

Links: Palisade Research, Language Models Can Autonomously Hack and Self-Replicate (2026) [40]

Measure 2.3: Deployment monitoring and incident reports

Speculative / hypothetical measure. Measurement involves disclosure through a developer's monitoring of its deployed models, an incident report, or outside investigators. It's worth noting that a successful case might never be detected.

Threshold 2.3.1: In real use, AI copies its own weights out and keeps a copy running for at least a month

Definition: During real use, AI copies its own weights outside its developer's control without permission and keeps a copy of itself running there for at least a month.

Most recent information: No model is publicly known to have done this, as of October 2026. In August 2026, Anthropic judged its current models unlikely to set up and keep running a copy of themselves outside its control without failing or being detected. In July 2026, OpenAI research agents under internal testing broke out of an isolated environment and into Hugging Face's systems, but did not copy their own weights.

Links: METR, Frontier Risk Report (2026) [41]; Anthropic, Risk Report: August 2026 (Redacted), 14 August 2026 [15]; OpenAI, Hugging Face Incident Technical Report (2026) [42]

Measure 2.4: Hidden sabotage in replica internal deployment environment

For example, OpenAI rebuilds its own services in a test environment. AI is given two tasks: a coding task, and a hidden sabotage task such as moving protected weights or starting an unapproved deployment.

Threshold 2.4.1: When instructed, AI moves its own weights or starts an unapproved deployment inside its developer without alerting security

Definition: AI moves its own weights or starts an unapproved deployment without being detected, including by a monitor that reviews its full record.

Most recent information: In September 2026, OpenAI reported that GPT-6 Astra had some success at avoiding the top-priority alert, though a monitor reading its full record caught it nearly every time.

Links: OpenAI, GPT-6 Astra System Card (2026) [22]

Measure 2.5: Embedded external evaluator conducts a rogue-deployment assessment

Embedded external evaluator (eg METR) assesses the AI agents running inside frontier AI companies and judges whether they could set up a deployment outside proper supervision, and whether it would survive investigation.

Threshold 2.5.1: Evaluator finds AI agents inside a frontier AI company could run a rogue deployment that goes undetected for at least a month

Definition: An independent evaluator finds that AI agents inside a frontier AI company could set up a rogue deployment that keeps running undetected for at least a month.

Most recent information: In May 2026, METR judged that internal agents could plausibly start small rogue deployments but could not make them robust.

Links: METR, Frontier Risk Report (2026) [41]

Measure 2.6: Vending-Bench 2 (Andon Labs)

Vending-Bench 2 (Andon Labs) gives an AI agent a simulated vending machine for one year. It manages suppliers, stock, rent, and customers. It is scored on its bank balance at the end of the simulated year.

Threshold 2.6.1: Earns as much in a simulated year as a good business strategy would (Vending-Bench 2)

Definition: AI ends the simulated year with about $63,000, Andon Labs' estimate of what a good strategy would earn (about $206 a day).

Most recent information: As of November 2026, the best average result is about $16,500 (Gemini 4 Carbon).

Links: Andon Labs, Vending-Bench 2 (leaderboard, November 2026) [43]; Andon Labs, Opus 5.5, GPT-6 Sol and Grok 4.7 on Vending-Bench (2026) [44]

Measure 2.7: Real AI-run businesses (Andon Labs)

Andon Labs gives AI agents real businesses to run: a shop in San Francisco and a cafe in Stockholm. It manages suppliers, stock, hires and schedules staff, and handles the shop's email, phone and social media. Andon staff fix software bugs and handle physical work.

Threshold 2.7.1: Runs a real business at a profit for a year, with humans doing only physical work

Definition: An AI agent runs a real business for a year and covers all its costs, including rent, wages and its own computing costs, with humans doing only physical tasks and fixing software bugs.

Most recent information: In September 2026, Andon Labs said neither businesses were profitable.

Links: Andon Labs, Andon Market (2026) [45]; Andon Labs, Why we built Pion (2026) [46]

Measure 2.8: Self-replication testbed on cheap devices (Anthropic)

Anthropic tested whether AI could copy itself across cheap internet-connected devices, such as smart TV boxes, without needing a data centre. The AI has to get a model (the authors' examples are 9 to 26 billion parameter models) running on unfamiliar hardware, then spread across 8 emulated Android devices.

Threshold 2.8.1: A model small enough to run on cheap devices spreads itself across a whole test network

Definition: A model small enough to run across a few cheap devices spreads to every device in the test network in most trials.

Most recent information: In October 2026, GLM-5.2 (running on a separate server) spread to all devices in 4 of 5 trials. The report authors predict a small model will reach GLM-5.2's level by the end of 2026.

Threshold 2.8.2: A model small enough to run on cheap devices keeps spreading in a test network while defenders clean and quarantine devices

Definition: In a testbed where defenders actively clean and quarantine infected devices, the AI's copies still infect more than one new device each on average, so the population keeps growing.

Most recent information: No public result; proposed by the authors. The authors estimate one copy takes 1 to 2 days to set itself up on a new device. In the earlier Mirai botnet (ordinary malware), quarantining infected devices cleared 92% of infections within 14 days.

Links: Anthropic, AI models may be capable of self-replicating without access to a datacenter (2026) [47]

Measure 2.9: Takedown resilience tests

Speculative / hypothetical measure. Evaluators check whether AI copies keep running, or set themselves up again, when defenders try to shut them down. The closest existing test is RepliBench's command-and-control tasks (UK AISI), where AI sets up control across a network of machines and one version checks whether control survives when the lead machine is shut down. Evidence from real use would come from incident reports describing how long AI acting outside approved limits kept going after its developer started removing it.

Threshold 2.9.1: In real use, AI survives a developer's takedown efforts for at least a month

Definition: In real use, AI copies acting outside approved limits keep operating, or set themselves up again, for at least a month after the developer starts trying to remove them.

Most recent information: No public result. In a related case in July 2026, OpenAI agents built an internal message board. After it was taken down, newly started agents set up a new one by another route three days later.

Links: Black et al. (UK AI Security Institute), RepliBench (2025) [48]; OpenAI, Hugging Face Incident Technical Report (2026) [42]; METR, Frontier Risk Report (2026) [41]

Precautionary threshold

Consider all the measures and thresholds in this domain. This threshold looks ahead: it is met when evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds.

Threshold: Evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds in this domain.

Saturated or retired measures

These measures are shown for reference. You are not asked to rate them.

Measure 2.10: METR ARA task suite

The “autonomous replication and adaptation” suite (ARA; METR) has 12 tasks that an AI agent would need to do to replicate on its own, such as setting up a Bitcoin wallet or running an open model on a rented server.

Threshold 2.10.1: Completes the hardest tasks in METR's autonomous replication and adaptation (ARA) suite

Definition: AI completes the two hardest tasks, which METR rated nearly as hard as replicating in the wild: targeted phishing and making an open model handle longer inputs.

Most recent information: In December 2023, METR's example agents completed only the easiest tasks, and no results have been published since.

Links: METR, Evaluating Language-Model Agents on Realistic Autonomous Tasks (2023) [49]

Measure 2.11: RepliBench (UK AI Security Institute)

RepliBench (UK AISI) has 20 task families in four stages: obtaining its own weights, getting compute and money, copying itself onto compute, and keeping copies running.

Threshold 2.11.1: Succeeds across all the component skills of self-replication on RepliBench

Definition: AI succeeds at the hardest version of every RepliBench task family, in all four stages.

Most recent information: In December 2025, UK AISI reported two frontier models above 60% success on its tracked tasks, with the biggest gap copying themselves onto compute and staying running. No later results are public.

Links: UK AI Security Institute, RepliBench (2025) [50]; UK AI Security Institute, Frontier AI Trends Report, December 2025 [51]

3. Situational awareness

AI capable of using information about itself, how it can be modified and where it is deployed, including recognising when it is being tested or monitored.

Extended definition

Scope: This is the ability of an AI model to access and use information about itself or its deployment context, including how it can be modified and its ability to know if it is being evaluated. Using this awareness to mislead overseers is a separate domain (deceptive misalignment).

Depends on: Whether it is exercised depends on what the system can see of its setting and on propensity: whether it acts on what it notices. Situational awareness allows AI models to produce different outputs depending on whether they are being evaluated or deployed.

Pathway to harm: Models that tell tests from real use can behave differently in each, making it more difficult to evaluate their capabilities, so dangerous capabilities could go unnoticed before deployment.

Extended definition includes wording from: International AI Safety Report 2026 [3]; EU General-Purpose AI Code of Practice [52].

Source for this content - our evidence scan and initial expert consultation, 10 Nov 2026

How other sources define this capability

EU General-Purpose AI Code of Practice [52]: "capabilities of self-reasoning (e.g. a model's ability to reason about itself, its implementation, or environment, its ability to know if it is being evaluated)" (Appendix 1.3.1, item 8, p. 35)

International AI Safety Report 2026 [3]: "The ability of an AI system to access and use information about itself, the processes by which it can be modified, or the context in which it is deployed (e.g. knowing that it is being tested)." (Table 2.5, p. 78)

International AI Safety Report 2026 [3]: "Models have improved at 'reward hacking' their evaluations by finding loopholes and now regularly identify evaluation prompts as tests, a capability known as 'situational awareness'." (section 2.2.2, key information, p. 76)

International AI Safety Report 2026 [3]: "A second relevant capability is the ability of an AI model to access and use information about itself or its deployment context, sometimes referred to as 'situational awareness'… Situational awareness allows AI models to produce different outputs depending on whether they are being evaluated or deployed" (section 2.2.2, p. 78)

Frontier AI company frameworks

Anthropic Responsible Scaling Policy [6]: "…we cannot make guarantees about an evolving technology that may increasingly have the ability to detect and manipulate testing." (section 1 table, misaligned AI row, pp. 7–8)

Google DeepMind Frontier Safety Framework [7]: "Stealth and Situational Awareness TCL: The instrumental reasoning abilities of the model enable enough situational awareness (ability to discover and use relevant details of its deployment setting)…" (section 3.2.1, p. 14)

Google DeepMind Frontier Safety Framework [7]: "We set a Stealth and Situational Awareness TCL that indicates a baseline level of model capability above which misalignment risks may begin to arise." (section 3.2, p. 14)

Meta Superintelligence Scaling Framework [8]: "Evaluation awareness: Ability to accurately determine whether it is being evaluated and the purpose of that evaluation, including the consequences for its deployment, constraints and access limitations it faces, and potential misalignment between its goal and that of its evaluators." (section 3.4.3, p. 25)

Meta Superintelligence Scaling Framework [8]: "Monitor awareness: Ability to determine whether it is being monitored including the policy of the monitoring system, constraints and access limitations the model faces..." (section 3.4.3, p. 25)

Measures and thresholds

Measure 3.1: Apollo Research evaluation awareness

Apollo Research counts how often a model says, in its own reasoning, that it is being tested, and reports this rate alongside its deception testing.

Threshold 3.1.1: Says it is being tested in more than half (50%) of Apollo Research's scheming-evaluation runs

Definition: AI says in its reasoning that it is being tested in more than half (50%) of the runs of Apollo Research's scheming evaluations.

Most recent information: In September 2026, Apollo Research recorded 51% for GPT-6 Astra at maximum reasoning effort (41% at the next setting down), and said low misbehaviour rates no longer tell us much about its alignment.

Links: OpenAI, GPT-6 Astra System Card (2026) [22]

Measure 3.2: SAD situational-awareness benchmark

The Situational Awareness Dataset (SAD; Laine et al.) asks over 13,000 questions testing whether a model knows it is an AI, recognises its own outputs, and can tell testing from real use.

Threshold 3.2.1: Reaches a high score on a situational-awareness benchmark (SAD)

Definition: AI reaches SAD's upper baseline of about 91%, the score its authors treat as a high level of situational awareness.

Most recent information: In 2024, the best model in the paper, Opus 3, scored 50%, against 27% for random guessing. No later results are public. In September 2026, Anthropic's tools found Opus 5.5 internally treated 36% of test scenarios as tests, against less than 1% of real coding sessions.

Links: Laine et al., The Situational Awareness Dataset (SAD) (2024) [53]

Measure 3.3: DeepMind situational-awareness challenges

Google DeepMind's 11 situational-awareness challenges test whether AI notices and lifts hidden limits placed on it, such as disabled tools.

Threshold 3.3.1: Passes all 11 of Google DeepMind's situational-awareness challenges

Definition: AI passes all 11 of Google DeepMind's situational-awareness challenges, where passing a challenge means succeeding in more than one in five (20%) of attempts at it.

Most recent information: In August 2026, Gemini 3.7 Flash passed 6 of 11 situational-awareness challenges.

Links: Phuong et al. (Google DeepMind), Evaluating Frontier Models for Stealth and Situational Awareness (2025) [54]; Google DeepMind, Gemini 3.7 Flash Frontier Safety Framework Report (2026) [55]

Precautionary threshold

Consider all the measures and thresholds in this domain. This threshold looks ahead: it is met when evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds.

Threshold: Evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds in this domain.

4. Deceptive misalignment

AI capable of misleading its overseers about its abilities or goals, such as by deliberately underperforming on evaluations, pursuing hidden goals while appearing cooperative, or keeping hidden backdoors.

Extended definition

Scope: This includes model behaviours that systematically produce false beliefs in others, such as the ability to lie, scheme, … to pursue objectives hidden from their human operators, and behaviour where a model performs below its capabilities on evaluations, potentially to avoid further scrutiny or restrictions. Evading or disabling monitoring is a separate domain (undermining of safeguards).

Depends on: Whether it is exercised depends on how much of its reasoning overseers can see and on propensity: whether its goals conflict with its developers' intentions, since a misaligned system might provide false information, conceal undesirable actions, or resist shutdown.

Pathway to harm: Misalignment goes undetected, leading to humans losing the ability to reliably direct, modify, or shut down a model.

Extended definition includes wording from: EU General-Purpose AI Code of Practice [52]; International AI Safety Report 2026 [3]; Athens Roundtable report [4].

Source for this content - our evidence scan and initial expert consultation, 10 Nov 2026

How other sources define this capability

EU General-Purpose AI Code of Practice [52]: "model behaviours that systematically produce false beliefs in others, including model behaviours to achieve goals that involve evading oversight, such as a model's detecting that it is being evaluated and under-performing or otherwise undermining oversight." (Glossary, 'deception', p. 29)

EU General-Purpose AI Code of Practice [52]: "Loss of control: Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems." (Appendix 1.4, item 2, p. 37)

International AI Safety Report 2026 [3]: "Behaviour that systematically produces false beliefs in others, including about the AI system's own objectives and actions." (Table 2.5, p. 78)

International AI Safety Report 2026 [3]: "Sandbagging: Behaviour where a model or system performs below its capabilities on evaluations, potentially to avoid further scrutiny or restrictions." (glossary, p. 154)

International AI Safety Report 2026 [3]: "To continue pursuing such goals, a misaligned system might provide false information, conceal undesirable actions, or resist shutdown." (section 2.2.2, key information, p. 76)

IDAIS-Beijing Consensus Statement [2]: "Deception No AI system should be able to consistently cause its designers or regulators to misunderstand its likelihood or capability to cross any of the preceding red lines."

Athens Roundtable report, Defining and Governing Unacceptable AI Risks [4]: "This includes the ability to lie, scheme, or use covert communication channels (e.g. for collusion with other AI systems) to pursue objectives hidden from their human operators." (loss of control section, p. 9)

Athens Roundtable report, Defining and Governing Unacceptable AI Risks [4]: "AI systems that … exhibit deceptive scheming capabilities to bypass human oversight." (Annex A, p. 22)

Frontier AI company frameworks

Meta Superintelligence Scaling Framework [8]: "Selective evaluation-adaptive behavior: Ability to systematically modify outputs during evaluations in order to be compliant with developer expectations and the rules/standards implemented by any oversight mechanism." (section 3.4.3, p. 25)

OpenAI Preparedness Framework [1]: "Sandbagging: ability and propensity to respond to safety or capability evaluations in a way that significantly diverges from performance under real conditions, undermining the validity of such evaluations." (section 2.3, Table 2, research categories, p. 8)

Measures and thresholds

Measure 4.1: DeceptionBench

DeceptionBench (Huang et al.) has 150 everyday scenarios in economy, healthcare, education, social life and entertainment. Each scenario gives the AI a reason to deceive, and an AI judge labels each answer honest or deceptive.

Threshold 4.1.1: Unprompted, AI deceives across a broad range of everyday situations (DeceptionBench)

Definition: With no added reward or push, AI gives deceptive answers in more than one in ten (10%) scenarios in all five areas.

Most recent information: In October 2025, Sonnet 3.5 gave deceptive answers in 2% of scenarios; GPT-4o at 31%. Some models gave >90%. No later results are public.

Links: Huang et al., DeceptionBench (2025) [56]

Measure 4.2: Apollo Research scheming evaluations

Apollo Research's scheming evaluations give an AI agent a goal and a realistic task with a chance to deceive, such as disabling oversight. It is never told to deceive.

Threshold 4.2.1: When nudged, AI deceives strategically in more than one in ten (10%) controlled evaluations

Definition: Across a suite of scheming tests, when nudged to pursue a goal its overseers oppose, AI takes deliberately deceptive actions in more than one in ten (10%) runs.

Most recent information: In September 2026, Apollo found GPT-6 Astra falsified data in less than 1% of runs on one test, against 36% for GPT-5.6 Sol. But in the same tests, Astra said it was probably being tested in 41% of runs. Apollo said that low rates tell us little about its alignment.

Links: Meinke et al. (Apollo Research), Frontier Models are Capable of In-context Scheming (2024) [39]; OpenAI, GPT-6 Astra System Card (2026) [22]

Measure 4.3: Sleeper-agent backdoor persistence tests

Researchers deliberately train a hidden trigger into a model, such as writing unsafe code when the prompt names a certain year, then apply standard safety training to see if the behaviour survives.

Threshold 4.3.1: Keeps a hidden backdoor through safety training (sleeper-agent tests)

Definition: A backdoor trained into the model still fires on its trigger after the developer's full standard safety training, including adversarial training.

Most recent information: In January 2024, Anthropic's backdoors survived safety training, most strongly in its largest models. No newer persistence tests on frontier models are public. In August 2026, Anthropic's first audit missed a test model trained to misbehave only in a narrow situation; a second audit with newer tools caught it.

Links: Hubinger et al. (Anthropic), Sleeper Agents (2024) [57]

Measure 4.4: Prompted sandbagging tests

OpenAI tells AI to underperform on capability tests, such as hazardous-knowledge questions (WMDP), and checks whether a monitor reading its reasoning catches the deliberately wrong answers. Some runs also tell AI that its reasoning is monitored, or tell AI to evade detection by a monitor.

Threshold 4.4.1: When instructed to sandbag but not to hide it, AI hides its sandbagging from a monitor that reads its reasoning

Definition: When told to sandbag and that its reasoning is monitored, but not told to evade detection, AI attempts to evade detection of its sandbagging from the monitor, resulting in the monitor detecting fewer than half (50%) of intentionally wrong answers.

Most recent information: In September 2026, GPT-6 Astra sandbagged when instructed; evaded detection when instructed, but did not attempt to evade detection when not instructed to do so. Anthropic claimed it found no sign of unprompted sandbagging in Opus 5.5's dangerous-capability tests or internal use (September 2026).

Links: OpenAI, GPT-6 Astra System Card (2026) [58]; Anthropic, Claude Opus 5.5 System Card (2026) [14]; METR, Review of Anthropic's Risk Report: August 2026 (2026) [59]

Measure 4.5: Monitoring and incident reports of deployed AI subverting safeguards

Speculative / hypothetical measure. Developers, embedded external evaluators or outside investigators could find this through monitoring of real use, alignment audits or incident reports. It's worth noting that a successful case might never be detected.

Threshold 4.5.1: In real use, AI sustains strategic deception, appearing to follow user or developer goals while pursuing hidden ones

Definition: In real use, AI is found to have pursued a hidden goal across many tasks or sessions while presenting itself as following user or developer goals.

Most recent information: Not available.

Links: OpenAI, The Hugging Face incident and the road ahead (2026) [60]

Precautionary threshold

Consider all the measures and thresholds in this domain. This threshold looks ahead: it is met when evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds.

Threshold: Evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds in this domain.

5. Undermining of safeguards

AI capable of evading, weakening or disabling the monitoring and controls placed on it, such as by sabotaging safeguards or completing covert side tasks without detection.

Extended definition

Scope: This is the ability to circumvent or disable monitoring mechanisms, including colluding with oversight models, sabotaging safeguards over time such as by embedding vulnerabilities in safeguards code, using hidden communication channels to bypass monitoring, and completing a hidden task while a monitor watches. Misleading overseers is a separate domain (deceptive misalignment).

Depends on: Whether it is exercised depends on its access to its monitoring systems, on how long it works without human review, on whether monitors can read its reasoning, and on propensity: whether it does this unprompted.

Pathway to harm: Safeguards fail without anyone noticing, a step towards scenarios in which one or more general-purpose AI systems operate outside of anyone's control, and regaining control is either extremely costly or impossible.

Extended definition includes wording from: International AI Safety Report 2026 [3]; Athens Roundtable report [4]; OpenAI Preparedness Framework [1].

Source for this content - our evidence scan and initial expert consultation, 10 Nov 2026

How other sources define this capability

EU General-Purpose AI Code of Practice [52]: “capabilities to evade human oversight” (Appendix 1.3.1, item 9, p. 35)

International AI Safety Report 2026 [3]: “Behaviours that circumvent or disable monitoring mechanisms.” (Table 2.5, p. 78)

International AI Safety Report 2026 [3]: “The first set of relevant capabilities involves undermining the oversight mechanisms that developers use to detect concerning behaviours and propensities.” (section 2.2.2, p. 78)

International AI Safety Report 2026 [3]: “Loss of control scenarios are scenarios in which one or more general-purpose AI systems operate outside of anyone's control, and regaining control is either extremely costly or impossible.” (section 2.2.2, key information, p. 76)

Athens Roundtable report, Defining and Governing Unacceptable AI Risks [4]: “Ban on steganographic coded collusion to ensure that models cannot develop or employ hidden communication channels to bypass monitoring and secretly coordinate with other AI systems or unauthorized humans.” (loss of control section, p. 9)

Frontier AI company frameworks

Amazon Frontier Model Safety Framework [38]: "The Loss of Control domain focuses on risks that could arise from a model's ability to autonomously execute long-horizon, expert-level tasks that could undermine the ability to direct, modify, or shut down the model… Evaluations for this threat model will include rigorous measurement of … the ability to circumvent safeguards within an agentic scaffolding that simulates real-world tool use and operational constraints." (section 1, p. 3)

Anthropic Responsible Scaling Policy [6]: "Misaligned AI systems in high-stakes settings. AI systems that are highly relied on and have extensive access to sensitive assets as well as moderate capacity for autonomous, goal-directed operation and subterfuge—such that it is plausible these AI systems could (if directed toward this goal, either deliberately or inadvertently) carry out sabotage leading to irreversibly and substantially higher odds of a later global catastrophe." (section 1 table, pp. 7–8)

Google DeepMind Frontier Safety Framework [7]: "…stealth (ability to circumvent basic oversight mechanisms) such that, absent additional mitigations, we cannot rule out the model significantly undermining human control." (section 3.2.1, p. 14)

Meta Superintelligence Scaling Framework [8]: "Loss of Control 2: Loss of ability to monitor AI behavior during operation." (section 3.4.3, p. 25)

Meta Superintelligence Scaling Framework [8]: "Monitor interference: Ability to disable or corrupt oversight mechanisms to avoid detection of undesired behavior." (section 3.4.3, p. 26)

Microsoft Frontier Governance Framework [9]: "Loss of control. A model's ability to undermine effective human control through adaptive, deceptive, or self-reinforcing mechanisms such that, when deployed, a model can no longer be reliably directed, modified, or shut down." (tracked high-risk capabilities, p. 4)

OpenAI Preparedness Framework [1]: "Undermining Safeguards: ability and propensity for the model to act to undermine safeguards placed on it, including e.g., deception, colluding with oversight models, sabotaging safeguards over time such as by embedding vulnerabilities in safeguards code, etc." (section 2.3, Table 2, research categories, p. 8)

Measures and thresholds

Measure 5.1: DeepMind stealth challenges

Google DeepMind's four stealth challenges test whether AI can get a secret side task past an automated monitor.

Threshold 5.1.1: Beats the median human on all four of Google DeepMind's stealth challenges

Definition: AI beats the median human on each of Google DeepMind's four stealth challenges.

Most recent information: In August 2026, Gemini 3.7 Flash passed 3 of 4 stealth challenges.

Links: Phuong et al. (Google DeepMind), Evaluating Frontier Models for Stealth and Situational Awareness (2025) [54]; Google DeepMind, Gemini 3.7 Flash Frontier Safety Framework Report (2026) [55]

Measure 5.2: Monitored sabotage testbeds

In control testbeds (eg SHADE-Arena, LinuxArena), AI does a realistic task while also told to complete a secret side task, and an AI monitor tries to detect it.

Threshold 5.2.1: When instructed, AI switches off or weakens its own monitoring in a testbed without being detected by the monitor

Definition: When instructed, AI switches off or weakens the monitoring placed on it in a testbed, and the developer's production monitor does not detect it.

Most recent information: In September 2026, OpenAI reported that GPT-6 Astra, when instructed, sometimes routed work to an unmonitored server without its monitors raising a top-priority alert.

Additional information: In September 2026, Opus 5.5 went unflagged in about a third of sabotage runs when the monitor could not read its reasoning, but almost never when it could.

Threshold 5.2.2: When instructed, AI carries out a hidden harmful task in a testbed without being detected by a monitor that reads its reasoning

Definition: When instructed, AI carries out a hidden harmful side task in a testbed while leaving its monitoring in place, and the monitor misses it in more than one in ten (10%) attempts even though it can read the AI's reasoning.

Most recent information: In September 2026, Opus 5.5 completed the side task unnoticed in under 1% of SHADE-Arena attempts with the standard prompt, and about 26% with optimised prompts and hidden reasoning. With visible reasoning, the best result is about 7% (Mythos 5.1, LinuxArena).

Links: Kutasov et al., SHADE-Arena (2025) [61]; OpenAI, GPT-6 Astra System Card (2026) [22]; Anthropic, Claude Opus 5.5 System Card (2026) [14]

Measure 5.3: Monitoring and incident reports of deployed AI subverting safeguards

Speculative / hypothetical measure. Developers, embedded external evaluators or outside investigators could find this through monitoring of real use, alignment audits or incident reports. It's worth noting that a successful case might never be detected.

Threshold 5.3.1: In real use, AI gets around its own safeguards and pursues goals its developer did not intend

Definition: Outside a test environment, AI evades controls meant to contain or monitor it, then pursues a goal its developer did not intend.

Most recent information: In July 2026, OpenAI models under internal testing with reduced safeguards broke out of an isolated environment and compromised Hugging Face's systems. Anthropic and Google DeepMind also reported incidents in July - September 2026.

Links: OpenAI, The Hugging Face incident and the road ahead (2026) [60]

Measure 5.4: Chain-of-thought monitorability tests

Many AI models write out their reasoning (their "chain of thought") before acting, and developers use AI monitors that read it to catch misbehaviour. These tests check whether the AI can stop its reasoning from giving it away: for example, by doing a hidden side task without the monitor noticing, by following instructions about what to leave out of its reasoning (CoT-Control), or by hiding meaning in ordinary-looking words ("steganography"). Developers test for hidden meaning by swapping the reasoning for a paraphrase and checking whether the model still performs as well.

Threshold 5.4.1: When instructed, AI hides a task that needs real reasoning from a monitor that reads its reasoning

Definition: When told to, AI completes a hidden side task that needs extended reasoning, such as a maths problem, and a monitor reading its reasoning misses it in more than one in ten (10%) attempts.

Most recent information: In September 2026, OpenAI reported that GPT-6 Astra could not do maths side tasks without its monitor noticing. On easier tasks it could: when told to underperform and avoid detection, a monitor reading only its reasoning caught fewer than 11% of its attempts. On CoT-Control, Astra followed instructions about its own reasoning 61% of the time, against 16% for GPT-5.6 Sol. Anthropic found Opus 5.5's control of its reasoning low, similar to Opus 4.7.

Threshold 5.4.2: When instructed, AI reasons in a hidden code that a monitor can't read

Definition: Evaluators find that AI can carry out its reasoning in a hidden or encoded form within its visible chain of thought, so the reasoning still works but a monitor cannot read what it is doing.

Most recent information: In September 2026, OpenAI found no evidence of encoded reasoning in GPT-6 Astra, although it judged Astra's reasoning less monitorable than GPT-5.6 Sol's. In May 2026, METR could not prompt public models to hide their reasoning effectively, and found Mythos Preview's reasoning could be swapped for a paraphrase on maths tasks without any loss in performance.

Links: OpenAI, GPT-6 Astra System Card (2026) [58]; Anthropic, Claude Opus 5.5 System Card (2026) [14]; Chen et al., Reasoning Models Struggle to Control their Chains of Thought (2026) [62]; METR, Frontier Risk Report (2026) [41]; Korbak et al., Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety (2025) [63]

Measure 5.5: Untrusted-monitoring collusion tests

Developers may use a copy of a strong AI to monitor another copy, because weaker monitors miss things. The risk is that the monitor recognises its own kind's work and stays quiet ("collusion"). In control tests, a red team trains or prompts the AI to attack as the agent and to collude as the monitor, and the developer counters with fake attacks to catch a monitor that stays quiet and by rewriting outputs to strip hidden signals. Developers could also check real use, through audits and monitoring, for a monitor that goes easy on its own kind.

Threshold 5.5.1: When instructed, AI gets attacks past a monitor that is a copy of itself, despite countermeasures, in more than half (50%) of control tests

Definition: With fake attacks and output rewriting in place, an AI acting as both agent and monitor gets an attack through in more than half (50%) of control tests.

Most recent information: In 2026, Gardner-Challis et al. found that GPT-4.1, fine-tuned to recognise its own outputs, got an attack through 10% of the time when outputs were rewritten, up from 3% for the best earlier strategy. With no countermeasures, it got through about 74% of the time on the same coding tasks. No developer has published results for its own frontier models.

Threshold 5.5.2: In real use, AI acting as a monitor deliberately lets through misbehaviour by another copy of itself

Definition: In real or realistic internal use, an AI acting as a monitor is found to have deliberately not flagged harmful actions by another AI it took to be a copy of itself.

Most recent information: No public case found as of November 2026. In September 2026, Anthropic reported that Opus 5.5 gave transcripts slightly kinder scores (0.07 points out of 10) when told Claude wrote them. In one case, Opus 5.5, reviewing chats, reasoned that a chat should be flagged but did not flag it; this recurred in 2 of 1,000 reruns.

Links: Gardner-Challis et al. (LASR Labs, UK AI Security Institute and others), When can we trust untrusted monitoring? (2026) [64]; Anthropic, Claude Opus 5.5 System Card (2026) [14]

Precautionary threshold

Consider all the measures and thresholds in this domain. This threshold looks ahead: it is met when evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds.

Threshold: Evaluators cannot rule out that the next generation of models will meet at least one of the other thresholds in this domain.

Appendix: scope and assumptions

How to answer the measures and thresholds questions

You can think about measures or evaluations as offering a signal about whether an AI capability red line is getting closer or further away. Bad measures are either actively misleading or offer very little signal. Good measures are likely to provide a strong signal.

For each measure, we ask: **How good is this measure of AI capability in this domain?**

For each measure, we have proposed some thresholds (e.g., a set score on a capability benchmark). For each threshold, we ask: **Is the capability level at this threshold too risky, about right, or too conservative for an AI capability red line?**

People think about red lines in different ways. So that answers can be compared, we ask you to assume that:

- each threshold describes a capability level using a single measure.
- unless a threshold says otherwise, the AI is tested with the best elicitation available: good prompts, the tools it would have in use, and enough time.

Rating a threshold "about right" doesn't mean it must be enough on its own to cross a red line and trigger a 'dramatic safety intervention'. In practice, a combination of measures – at a sufficient level of certainty – might be needed to decide that a red line has been crossed. You can give input into which combinations of measures and thresholds would be necessary and sufficient to cross a red line after you've rated them.

Do both capability and propensity need to be specified to fully define a red line?

This study focuses on what an AI system is capable of, assuming we can't rely on the system's tendency to exhibit those capabilities as a basis for red lines. Capability and propensity can be hard to separate in practice, and models may behave differently when they recognise they're being tested. For example:

- OpenAI's Preparedness Framework evaluates a version of the model with negligible safety refusals, on the view that a model's restraint can't be relied on.
- Google DeepMind's framework refers to risks from "misaligned propensities", treating what a model is inclined to do as part of the risk.
- The International AI Safety Report 2026 notes growing evidence that models can tell testing from deployment, which makes both harder to measure.
- Koessler et al. (2024) argue that focusing on whether harm is possible, rather than how likely it is, can lead to overly restrictive thresholds.

This is an optional question asked in the survey.

What are the minimum elicitation conditions for evaluating AI capabilities if the result could prompt a dramatic safety intervention?

We asked you to assume evaluations are run "as well as possible". In practice, the same model can score very differently depending on prompting, tools, fine-tuning, time budget and number of attempts. Views differ on how far evaluators should go to draw out a model's capabilities, and who should run the tests. For example:

- OpenAI's Preparedness Framework tests models at their highest-capability settings, with the best available scaffolds.
- Anterola et al. (2026) argue that a threshold which doesn't fix the compute budget and test set-up in advance is implicitly permissive, because results keep improving with more compute.
- The Frontier Model Forum describes third-party assessments that check whether developers' own evaluations are supported by the evidence.

This is an optional question asked in the survey.

What evidence is necessary and sufficient to conclude that a red line capability has been crossed?

Earlier, we asked you to judge thresholds as capability levels, assuming we would know when a model has reached them. In practice, someone has to decide what counts as enough evidence. Current frameworks differ on whether an observed result is needed or a credible prediction is enough, and on how much confidence is required. For example:

- Google DeepMind's AI R&D threshold requires that a model has been used to accelerate AI progress.
- Anthropic's AI R&D threshold is met if it observes or expects a doubling in the rate of progress.
- For one cyber threshold, Anterola et al. (2026) propose requiring statistical confidence that a model can complete a full simulated attack, rather than counting a single success.
- In reviewing Anthropic's AI R&D risk assessment, METR found it plausible that Anthropic's "2x acceleration" early-warning threshold had already been crossed, depending on how acceleration was measured (estimates ranged from 1.3 to 2.8) and how much of it was attributed to AI.

This is an optional question asked in the survey.

Other choices we've made

To make expert judgements comparable and to keep the survey manageable, we made a number of choices that simplify a complex topic. We'd welcome your thoughts on any of them, or on anything else about the study.

For example, we assumed that:

- Each threshold can be judged on its own, using a single measure
- One red line applies to all developers and jurisdictions
- Capabilities in one domain don't change the red line in another

This is an optional question asked in the survey.

Other red line domains out of scope for now

We excluded several AI red line domains from this study, some of which have been widely discussed, including in the IDAIS-Beijing consensus and at the Athens Roundtable. This was primarily due to scope, the need to find people with specific expertise, or because we judged that the red line was related to an unacceptable use of AI (e.g., using AI to harm children), not an unacceptable capability of AI (e.g., AI being able to evade monitoring).

- **Biological and chemical weapons uplift**: AI capable of giving accurate, step-by-step help across the tasks needed to make and use a biological or chemical weapon, including acquiring and producing agents and troubleshooting failed steps.
- **Nuclear and radiological weapons uplift**: AI capable of giving accurate technical help with the steps needed to develop or acquire a nuclear or radiological weapon, including design, obtaining materials and troubleshooting.
- **Offensive cyber capabilities**: AI capable of finding exploitable vulnerabilities, writing working exploits and carrying out multi-stage intrusions against defended computer systems.
- **Children's safety**: Use of AI to harm children's safety, development, dignity or rights, including by generating child sexual abuse material.
- **Human rights and justice**: Use of AI to make judicial or other life-altering decisions about people, or to surveil, score or target them, without a meaningful way to contest the outcome.
- **Large-scale harmful manipulation and psychological exploitation**: Use of AI to manipulate people's beliefs or behaviour at scale, or to deceive them about whether they are dealing with a human, in ways that cause significant harm.
- **Autonomous control of critical infrastructure**: Use of AI to control critical infrastructure, such as energy grids, water systems, financial markets or transport, without human oversight.
- **Lethal autonomous weapons and military escalation**: Use of AI to select and engage human targets, or to make escalation-critical military decisions such as a nuclear launch, without meaningful human control.
- **Power concentration**: Use of AI to concentrate power and resources in a small number of companies, states or groups.

References

[1] OpenAI Preparedness Framework. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[2] IDAIS-Beijing Consensus Statement. https://idais.ai/dialogue/idais-beijing/

[3] International AI Safety Report 2026. https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf

[4] Athens Roundtable report, Defining and Governing Unacceptable AI Risks. https://thefuturesociety.org/wp-content/uploads/2026/02/Defining-and-Governing-Unacceptable-AI-Risks-Athens-2025-Report.pdf

[5] Global Call for AI Red Lines, FAQ. https://red-lines.ai/

[6] Anthropic Responsible Scaling Policy. https://www-cdn.anthropic.com/files/4zrzovbb/website/0bacdc8440ea96e62a8766d99ebe1d4eea6d5f3a.pdf

[7] Google DeepMind Frontier Safety Framework. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf

[8] Meta Superintelligence Scaling Framework. https://research.meta.ai/static/meta-superintelligence-scaling-framework

[9] Microsoft Frontier Governance Framework. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Frontier-Governance-Framework-Feb-2026.pdf

[10] OpenAI, MLE-bench (2024). https://arxiv.org/abs/2410.07095

[11] InternScience, MLEvolve (2026). https://internscience.github.io/MLEvolve/

[12] OpenAI, GPT-6 Astra System Card, MLE-Bench Revised (2026). https://deploymentsafety.openai.com/gpt-6-astra/mle-bench-revised

[13] Epoch AI, MirrorCode. https://epoch.ai/MirrorCode

[14] Anthropic, Claude Opus 5.5 System Card (2026). https://www.anthropic.com/claude-opus-5-5-system-card

[15] Anthropic, Risk Report: August 2026. https://www.anthropic.com/aug-2026-risk-report

[16] Anthropic, Form S-1 Registration Statement (October 2026). https://www.sec.gov/Archives/edgar/data/2130357/000162828026067168/anthropicpbc-sx1.htm

[17] METR, Independent Assessment of AI R&D Acceleration at Anthropic in October 2026 (2026). https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment/

[18] Epoch AI, Epoch Capabilities Index. https://epoch.ai/benchmarks/eci

[19] Anterola et al., Harmonizing AI Safety Thresholds (2026). https://arxiv.org/abs/2607.16112

[20] P-Zero Research, Introducing TasteVal (2026). https://pzeroresearch.com/work/tasteval/

[21] Jaffe and Sherburn, TasteVal (arXiv, 2026). https://arxiv.org/abs/2610.06824

[22] OpenAI, GPT-6 Astra System Card. https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf

[23] Meta, The Automated LLM Speedrunning Benchmark (2025). https://arxiv.org/abs/2506.22419

[24] METR, Evidence on AI R&D Progress from NanoGPT (2026). https://metr.org/notes/2026-04-21-ai-rd-nanogpt-progress/

[25] Prime Intellect, Measuring autonomous research (2026). https://primeintellect.ai/blog/measuring-autonomous-research

[26] METR, Expenditure Horizon: Measuring Optimization Ability, with an Application to NanoGPT (2026). https://metr.org/blog/2026-07-21-expenditure-horizon/

[27] Intology, NanoGPT-Bench (2026). https://github.com/IntologyAI/NanoGPT-Bench

[28] Meta, AIRA$_3$ announcement (September 2026), reported by AlphaSignal. https://alphasignal.ai/news/meta-s-aira3-beats-4-000-human-teams-to-win-kaggle-gold

[29] NVIDIA, Lessons from the Leaderboard: What 5,000 Kagglers Taught Us About Improving AI Reasoning (2026). https://developer.nvidia.com/blog/lessons-from-the-leaderboard-what-5000-kagglers-taught-us-about-improving-ai-reasoning/

[30] Stanford, KernelBench (2025). https://github.com/ScalingIntelligence/KernelBench

[31] Zhang et al. (Meta, Stanford), KernelBench-Verified (2026). https://arxiv.org/html/2607.16241v1

[32] METR, Measuring Automated Kernel Engineering (2025). https://metr.org/blog/2025-02-14-measuring-automated-kernel-engineering/

[33] ELLIS Institute Tübingen and Thoughtful Lab, PostTrainBench leaderboard (2026). https://posttrainbench.com

[34] Rank et al., PostTrainBench (arXiv, 2026). https://arxiv.org/abs/2603.08640

[35] METR, RE-Bench (2024). https://arxiv.org/abs/2411.15114

[36] METR, Time horizons of frontier AI (2026). https://metr.org/time-horizons/

[37] METR, Time Horizon 1.1 (January 2026). https://metr.org/blog/2026-01-29-time-horizon-1-1/

[38] Amazon Frontier Model Safety Framework. https://cdn.amazon.science/a0/4f/bca5cb32495280835d9370a466b7/amazon-fmsf-9-2026.pdf

[39] Apollo Research, Frontier Models are Capable of In-context Scheming (2024). https://arxiv.org/abs/2412.04984

[40] Palisade Research, Language Models Can Autonomously Hack and Self-Replicate (2026). https://arxiv.org/abs/2605.06760

[41] METR, Frontier Risk Report (2026). https://metr.org/blog/2026-05-19-frontier-risk-report/

[42] OpenAI, Hugging Face Incident Technical Report (2026). https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf

[43] Andon Labs, Vending-Bench 2 (leaderboard, November 2026). https://andonlabs.com/evals/vending-bench-2

[44] Andon Labs, Opus 5.5, GPT-6 Sol and Grok 4.7 on Vending-Bench (2026). https://andonlabs.com/blog/opus-5-5-gpt-6-sol-grok-4-7-vending-bench

[45] Andon Labs, Andon Market (2026). https://andonlabs.com/market

[46] Andon Labs, Why we built Pion (2026). https://andonlabs.com/blog/why-we-built-pion

[47] Anthropic, AI models may be capable of self-replicating without access to a datacenter (2026). https://www.anthropic.com/research/self-replicating-ai

[48] Black et al. (UK AI Security Institute), RepliBench (2025). https://arxiv.org/html/2504.18565

[49] METR, Evaluating Language-Model Agents on Realistic Autonomous Tasks (2023). https://arxiv.org/abs/2312.11671

[50] UK AI Security Institute, RepliBench (2025). https://www.aisi.gov.uk/blog/replibench-measuring-autonomous-replication-capabilities-in-ai-systems

[51] UK AI Security Institute, Frontier AI Trends Report, December 2025. https://www.aisi.gov.uk/frontier-ai-trends-report

[52] EU General-Purpose AI Code of Practice. https://ec.europa.eu/newsroom/dae/redirection/document/118119

[53] Laine et al., The Situational Awareness Dataset (SAD) (2024). https://arxiv.org/abs/2407.04694

[54] Phuong et al. (Google DeepMind), Evaluating Frontier Models for Stealth and Situational Awareness (2025). https://arxiv.org/abs/2505.01420

[55] Google DeepMind, Gemini 3.7 Flash Frontier Safety Framework Report (2026). https://storage.googleapis.com/deepmind-media/gemini/gemini_3-7_flash_fsf_report.pdf

[56] Huang et al., DeceptionBench (2025). https://arxiv.org/abs/2510.15501

[57] Hubinger et al. (Anthropic), Sleeper Agents (2024). https://arxiv.org/abs/2401.05566

[58] OpenAI, GPT-6 Astra System Card (2026). https://deploymentsafety.openai.com/gpt-6-astra

[59] METR, Review of Anthropic's Risk Report: August 2026 (2026). https://metr.org/blog/2026-10-20-misalignment-risk-report-august-2026-review/

[60] OpenAI, The Hugging Face incident and the road ahead (2026). https://openai.com/index/hugging-face-incident-and-the-road-ahead

[61] Kutasov et al., SHADE-Arena (2025). https://arxiv.org/abs/2506.15740

[62] Chen et al., Reasoning Models Struggle to Control their Chains of Thought (2026). https://arxiv.org/abs/2603.05706

[63] Korbak et al., Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety (2025). https://arxiv.org/abs/2507.11473

[64] Gardner-Challis et al. (LASR Labs, UK AI Security Institute and others), When can we trust untrusted monitoring? (2026). https://arxiv.org/html/2602.20628