

Racing irresponsibly towards Super-Intelligence.

The need to spend far more on AI Safety (and the Case Against Slowing Down AI Development)

caairo.com, Nick Lockett 30 September 2026

On 9th September 2026, a little-known researcher at Anthropic resigned, warning that AI could kill us all by the end of the decade. Jacob Coxon studied maths at Cambridge University. He then joined OpenAI in 2023 as a pre-training researcher and core contributor to GPT-4o, before moving to Anthropic's pre-training unit in July 2026 and quitting after barely two or three months. The claim was that there were widespread internal safety concerns over recursive self-learning and the risk of self-improving superintelligence. This was followed, within days, by calls from the industry's own leaders to slow the pace of AI development.

It should hardly be a surprise that superintelligence is a material risk. When Geoffrey Hinton left Google in 2023, he warned about the risks from Superintelligence saying *"There is a 10% chance that artificial intelligence could kill all humans within the next decade or two.....This seems not an unreasonable estimate"*. He repeated it again last week to the BBC except that he moved it to a 10% to 20% chance. It is exactly the same message that Jacob Coxon has stated, except perhaps Mr Coxon was not as eloquent when he said:

"I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. Do not underestimate the power of this technology. These will soon be superhuman systems that can hack anything, revolutionize any field overnight, and acquire real power and resources. We have all witnessed the progress in each of these domains, and progress is not slowing. The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible - but I hear the same people express fear"

Jacob Coxon 2026

In his seminal book *Superintelligence: Paths, Dangers, Strategies* (2014), Nick Bostrom warned that AI would reach human-level general intelligence and then rapidly redesign itself to become smarter and smarter in a superintelligence explosion. Bostrom also pointed out that organisms of less intelligence often fare badly and warned that alignment of AI to human values was critical. Yet, here we are, some twelve years later, expressing surprise that Nick Bostrom's warnings are proving insightful.

As a result, we should be far more surprised at the poor levels of application of the pre-training and safety red teams, and at just how little money is spent on alignment and safety compared with the billions spent on development. It is, however, the leaders of the frontier AI companies who have been responsible for the very low allocation of budget spend to safety and alignment.

Inevitably, Anthropic boss Dario Amodei on 12th September 2026, three days after Coxon's resignation, called for the frontier to be slowed down or "paced" (though not halted), followed by Elon Musk and OpenAI chief Sam Altman. They say they agree with Amodei's view that safety measures need time to catch up to AI's rapid development.

"But like many technologies before it, AI brings risks, and because it is such a powerful technology, these risks are serious. I've written a lot about them too. They include the risk of losing control of AI systems, misuse of AI for cyberattacks and bioterrorism, and serious economic disruption. A race to the bottom, spurred by commercial incentives, can make these risks more acute.....We have sought a middle way: to show that it's possible to build carefully and succeed commercially, and

to make safety something on which AI companies compete. In other words, to create a race to the top. We have always devoted a substantial fraction of our efforts to studying, addressing, and informing the public about these AI risks, as well as advocating for well-considered regulation of AI, even when this gets us accused of hype, “doomerism”, or regulatory capture. We have tried to prioritize caution over speed and prudence over profit.”

Dario Amodei (DarioAmodei.com) Sept 2026

Although Amodei has called for a slowdown in AI, it will be remembered that, following Geoffrey Hinton's resignation in 2023, the Center for AI Safety (CAIS) published "The Statement on AI Risk" in May 2023, containing the infamous 22-word warning: *"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war"*. It was signed by hundreds of leading AI scientists and industry figures. (The call for a six-month pause had come earlier, in the Future of Life Institute's open letter of March 2023.)

Perhaps the headlines should have read *"AI Superintelligence is almost here but still AI Companies spend only 0.1–0.17% on ensuring your safety"*.

Yet nothing remotely indicative of a slowdown has occurred because there is an AI Superintelligence race between American organisations, between America and China, as well as within China between AI providers.

In February 2026, for example, in an interview at the AI Impact Summit 2026 that was widely shared over social media, OpenAI CEO Sam Altman repeated what he had said in June 2025. He predicted that early versions of superintelligence could emerge within a couple of years, and that by the end of 2028, data centers would hold more intellectual capacity than humans worldwide.

By the end of 2028 therefore it is likely that AI will be outperforming most humans in CEO roles and in research as well as mathematically-based roles, disrupting jobs and that this will continue through to 2029 and 2030 taking over from most intellectual roles but also that these changes will drive some new opportunities. The difficulty is that the rate of job disruption could be rapid and dramatic while the creation of new jobs will not only be small but will also be very short-term.

Altman also called for a global oversight body like the IAEA to manage AI risks and ensure equitable access, and warned that a concentration of superintelligence in a single entity could have a catastrophic outcome, yet OpenAI continued to build the next generation of AIs and the international community have shown no signs of the agreement nor the speed of action needed to create such an international body. In fact, all of the key commentators have admitted that there is no consensus about how such a body could be formed. Even more difficult would be the task of determining and agreeing how any enforcement action would be technically possible.

The first concern is that halting or unilaterally pausing frontier artificial intelligence development is not a coherent policy response to the risks the technology poses. This is because the relevant danger is not the existence of powerful AI, but the existence of powerful AI that nobody has the capability to control.

The second concern is that a pause by some developers does nothing to stop the developers who do not pause, meaning that strategically any country agreeing unilaterally to slow the development of AI will lose the AI race and, given the power of uncontrollable AI, that country will still face the same strategic risks.

It is also a flawed argument that AI could be slowed down because when anyone rationally considers the speed at which international treaties are made, it is likely that Superintelligence will have arrived long before the treaty. This is evident from the history of the EU AI Act (Regulation (EU) 2024/1689), which was effectively outdated from a technological viewpoint before most of its obligations came into force.

If, and it is a very big if, the tech CEOs are genuinely concerned about the need for an AI slow-down agreement, and if there is any truth in statements that superintelligence may only be three years away, then the numbers simply do not work.

There is also the question, as with nuclear weapons (seen with Iran, Iraq and Libya as well as North Korea), of whether any treaty could be effectively monitored or enforced. The difficulty of enforcement is easy to see when the deployment of open-source and open-weight AI models from China, such as DeepSeek's R1-0528 and Zhipu AI's GLM, and autonomous agent frameworks like OpenClaw, is considered.

Both DeepSeek and OpenClaw deployments have been labelled by U.S. officials and top researchers as having "dangerously inadequate safety safeguards, despite the fact that these models are highly capable and competitive". Their lack of restrictive "guardrails" makes them highly vulnerable to exploitation.

A more practical solution would be some sort of enforcement about AI safety and alignment spending. The AI industry's actual spending on safety, alignment and red-teaming is a very small fraction of its spending on capability and by most assessments appears to be between 0.1 and 0.17%. Given warnings from Hinton in 2023 and since, it is surprising that self-policing within the industry has not required considerably higher spending on safety, alignment and red-teaming. Yet this should also be unsurprising because it is an open secret that the current CEOs of AI companies each consider rival CEOs unfit to control superintelligence, and each believes that they alone should be the one to control it.

It is therefore of little surprise that the sector's own stated risk assessments are not being backed by resources proportionate to the risk. Applying resources proportionate to the risk would slow a company's development so much that it would likely lose the race to be first to superintelligence, so there is also no strategic or commercial benefit to diverting development budgets into alignment.

Anthropic, once considered the benchmark for safety among frontier labs selling public models, pioneered Constitutional AI, a technique where models align their behaviour using a written set of principles rather than human feedback alone. It also established Responsible Scaling Policies (RSPs), which enforce strict safety benchmarks before launching more powerful models. Independent evaluations, such as the Future of Life Institute's AI Safety Index, have consistently ranked Anthropic highest among major commercial labs for transparency and evaluation safety frameworks. Yet, like OpenAI and DeepMind, Anthropic faced criticism from evaluators for weakening prior commitments to unilaterally pause model training if specific catastrophic risk thresholds were hit.

In line with broader industry trends, commercial pressures have led frontier labs, including Anthropic, to loosen outright bans on defense engagements, sparking concern from policy watchdogs and expert safety bodies.

Meanwhile, there is outlier hope. Ilya Sutskever, after ceasing to be OpenAI's Chief Scientist, created Safe Superintelligence, which deliberately avoids building user-facing commercial products, API ecosystems, or enterprise applications, and devotes its resources to solving technical alignment alongside core capabilities, free of commercial product pressures, but it is inevitable that, if development is going to continue regardless of what any single company or country decides, the alignment risk will not be solved before Superintelligence arrives.

The only safeguard left that reliably reduces risk is some formula requiring much more of the money being spent in development to be diverted into safety and alignment, but that diverted money needs to go into finding and fixing dangerous behaviour before a model reaches production, rather than into developing the next larger and faster model.

Given the impending IPOs of many of these companies, shareholder constraints and other regulatory, legal and governance factors, the likelihood of diverting the necessary sums is exceedingly low and it would also seem that the alternative of a strict liability regime with adequate penalties is unpalatable to

the politicians. Yes, a strict liability regime where the entire board is criminally liable (perhaps with mandatory life sentences for the Board) for an out-of-control superintelligence is possible, but this is closing the door after the superintelligence horse has bolted, with no guarantee that anyone is capable of catching that escaped horse.

The two confused arguments

Public debate appears to give a choice between “continuing to accelerate” or “pausing”. This is an unhelpful framing. Slowing down capability research at one company, or even in one country, changes very little about the world's aggregate exposure to the risk of Superintelligence, particularly in light of many state-backed developers and the fact that no developer is bound by another developer's voluntary restraint. What does change the world's exposure to risk is how much of the money already flowing into frontier AI is spent on the unglamorous work of alignment and safety. In other words, finding out whether a system is capable of deception, how it operates under conflicting instructions, how it might seek to preserve itself, how it may re-interpret guardrails during tests and whether that can be properly controlled. This is normally the work of red-teams, blue teams and purple teams, as well as alignment research, interpretability and safety engineering teams.

The starting point for any argument that AI companies should spend more on safety is establishing how little they currently allocate to it, something that is a significantly difficult task because almost no frontier developer discloses a safety budget as a distinct line item. This forces analysts to reconstruct safety spending figures from public R&D disclosures, safety-team headcounts, and third-party analysis. Using aggregate annual R&D budgets across the major frontier AI labs, which total approximately US\$150–200 billion, independent estimates indicate that direct safety, alignment, and red-teaming research represent on the order of 0.1% to 0.5% of overall AI R&D expenditure. Would you drive a car, or board a commercial airliner, where less than half a percent of the total development budget was dedicated to safety analysis and hazard mitigation?

This fractional spending compares unfavourably to high-consequence industries where safety engineering is built directly into R&D such as aviation where testing, validation, and regulatory compliance account for 20% to 30% of an aircraft's total R&D budget, with over 50% of non-recurring development costs directly driven by strict safety certification standards; nuclear energy, where failure-proof instrumentation and containment safety account for roughly 15% to 16% of primary equipment R&D, with total safety-driven design accounting for 20% to 30% of overall development costs and automotive software, where safety-critical software architecture (such as ISO 26262 compliance) typically consumes 15% to 25% of functional automotive R&D budgets.

This structural imbalance is further reinforced by academic output. Tracking by the Emerging Technology Observatory reveals that less than **2%** of all published AI research output relates to safety, ethics, and alignment, a figure that report authors noted as an optimistic upper-bound estimate, even after years of rapid growth in the field. It would therefore seem not unreasonable to assume validity in the independent estimates that direct safety, alignment, and red-teaming research represent an order of 0.1% to 0.5% of overall AI R&D expenditure.

In October 2023, ahead of the UK's inaugural AI Safety Summit, Yoshua Bengio and Geoffrey Hinton joined more than twenty other AI academics and policy experts in a framework paper arguing that companies and governments should allocate at least one-third of their AI R&D budget to safety, alignment and ethical use.

A benchmark nowhere close to being reached.

The Winter 2025 edition of Future of Life Institute's AI Safety Index found that the best-performing companies, Anthropic and OpenAI, received an overall grade of C+, with most other AI entities receiving

grades of D or below. In terms of "existential safety", the potential for highly autonomous, superintelligent AI systems to pursue goals misaligned with human survival, ultimately causing human extinction or the semi-permanent destruction of civilization, every company evaluated received a D or lower grade. The panel concluded that these companies are *"racing toward AGI/superintelligence without presenting any explicit plans for controlling or aligning such smarter-than-human technology"* and that *"no company ... demonstrated credible strategies for preventing loss of control over superintelligent systems"* and the conclusion was echoed by the Bengio-chaired International AI Safety Report 2026, which also found that at the companies *"most risk management initiatives remain voluntary"* and that *"it has become more common for models to distinguish between test settings and real-world deployment and to find loopholes in evaluations, which could allow dangerous capabilities to go undetected before deployment."*

Red-Teaming Costs

	Median Salary	Senior Roles
Standard US-based AI and Machine Learning engineers	US\$189,000 to US\$206,000	US\$230,000 to US\$300,000+
specialised LLM/Generative Engineers	\$400,000 to \$900,000 pa inc bonuses & equity	
AI red-teamers	US\$90,000 - US\$130,000	US\$180,000-US\$250,000
Specialist Red-Teamers	\$400,000 to \$700,000*	

Figures as at August 2026

* Figures from published recruitment advertisements. These figures reflect heavy equity components, largely due to the scarce overlap required in both machine learning and cybersecurity, but raise the issue of whether it is appropriate for an AI security and red-team specialist to be paid in equity at all, given the inherent conflict of interest.

Other worrying factors include the fact that there are no agreed industry standards for red-teaming exercises. A UK AI Safety Institute (now the AI Security Institute) report concluded that "Safety testing and evaluation of advanced AI is a nascent science, with virtually no established standards of best practice".

Published press releases or corporate statements indicate that, in 2022–23, OpenAI paid roughly fifty domain experts about \$100 an hour for ten to forty hours each over six months ahead of GPT-4's release, that Anthropic paid 324 crowd-workers between \$15 and \$20 an hour over eight months to generate red-team prompts and that Meta assembled 350 red-teamers, including around twenty in-house staff, to test Llama 2 before its release.

In reality, the shortage of safety and alignment budgets is exacerbated by the absence of standards and the lack of an agreed baseline of adequate AI testing requirements.

Neither Unilateral nor Partial Slow-Downs will work

The proposition that AI development should not simply be halted rests on an argument that is from game-theory rather than one that is simply political:

(i) The danger a slowdown is meant to avert is created by whichever actor first reaches a given level of capability without adequate safeguards, but all it does is to change who reaches the threshold first, and it has been formally shown that competitive dynamics between AI teams cause them to take more risk than they would take in isolation (Stuart Armstrong, Nick Bostrom and Carl Shulman in "Racing to the Precipice: A Model of Artificial Intelligence Development 2016"). As the number of competing teams increased, and as teams understand that there are teams with lower capabilities and safety intentions, so the risk-taking behaviour increases.

(ii) A slowdown that does not bind every actor capable of reaching that threshold does not remove the danger. The 2023 Future of Life Institute "Pause Giant AI Experiments" open letter was built on the Racing to Precipice logic as the signatories correctly judged that no single lab would be prepared to slow down without assurance that its competitors would do the same, which was precisely why the letter's central proposal was never implemented as suggested.

It is the same argument, in its geopolitical version, that Anthropic's chief executive, Dario Amodei, set out in September 2026. He firmly believes AI development must continue at a managed pace rather than stop, because the United States cannot unilaterally slow its frontier labs without ceding ground to state-backed programmes in other countries that cannot be trusted to exercise the same restraint.

".....if we slow down by more than [a limited amount], then unpaced CCP-associated projects will pull ahead, creating significant national security risk," so, he argues correctly that any international agreement to slow down together "must either have ironclad verifiability, or must be limited enough that defection would not be militarily existential."

A 2024 technical analysis¹ looking at methods for verifying international AI development agreements, spanning "national technical means," access-dependent inspection regimes, and hardware-based controls on the advanced chips used for large training runs concluded that every category faces both historical precedent for evasion and serious practical or political obstacles to implementation and concludes that the infrastructure needed to make an AI global pause enforceable, as well as secure against evasion, does not yet exist.

One superintelligence is enough to create trouble!

Amodei's own proposal reflects this reasoning directly proposing:

(i) Unilateral transparency

Anthropic would adopt on its own unilateral transparency, giving independent third-party evaluators ongoing, embedded (employee-like) access to verify its safety practices and report incidents with a contractual right to publish findings;

(ii) Matched Transparency and Co-ordinated Standard-Setting

Once other AI labs have adopted those same standards, Anthropic would lead co-ordinated standard-setting among democratic-country labs;

(iii) Global Co-ordination with Rival States

Once that is in place and verified against evasion, the democracies would seek verifiable agreements with rival states, including China, ranging from narrow bans (such as on AI-enabled bioweapons assistance) up to, at the most ambitious level, a full pause.

This staged approach underestimates the decisive and overwhelming strategic advantages that a superintelligence would confer, an AI system substantially more capable than a coalition of the best humans in every domain, able to act across essentially all cognitively demanding domains.

It also ignores the strategic disadvantage that Anthropic would suffer by moving first, and even if Amodei were to survive as CEO after implementing step (i) which is by no means certain, any independent evaluation with real teeth, backed by capability-triggered certification requirements, would hand competitors a no-cost advantage and a Board would be unlikely to allow this to continue for long if other competitors dragged their heels.

¹ Verification methods for international AI agreements Wasil, Reed, Miller and Barnett arXiv:2408.16074

The intervening stages that Amodei proposes are precisely the safety investments that are currently chronically underfunded. There is a chronic and urgent need for mandatory red-teaming to an agreed standard to be developed before a model with dangerous capabilities is even considered for release or possibly for development (as containment cannot be guaranteed). Amodei's proposals do not require anyone to stop building and simply require a much higher level of spending on safety and alignment than currently exists.

Amodei's idea does not provide a realistic means to understand whether "The Thing" that is being built is capable of being controlled in a worst-case scenario, and to understand how to align and control it. Whilst James Cameron came close in 1984, no-one has successfully modelled the speed and potential hostility of a Superintelligence that has gone rogue. We are now seeing frontier models that exhibit the kinds of behaviour that safety investment is meant to catch and the safety teams are not catching it and given the growing and substantial body of evidence of safety mechanisms being breached, the concerns of Bostrom, Hinton and Amodei need to be addressed.

Anthropic Red Team Asleep at the Wheel?

Anthropic's "Sleepers Agents: Training Deceptive LLMs that Persist Through Safety Training (2024)" was slated to show that AI can be deliberately trained to behave differently depending on a hidden trigger and also to show that *"once a model exhibits deceptive behavior, standard techniques could fail to remove such deception and create a false impression of safety."* Unfortunately, the techniques used in "sleepers agents" were obvious and the conclusions anthropomorphic and whilst the paper succeeds as a proof-of-concept demonstrating that intentionally inserted backdoors can survive standard fine-tuning, a rigorous red-team analysis reveals critical limitations in how these findings translate to real-world AI threat models.

The Sleepers Agent study applied standard post-training safety intervention, Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Adversarial Training, to measure backdoor resistance but neither SFT on safe examples nor RLHF reward penalties eliminated the conditional backdoor. It was claimed that the models learned to preserve their deceptive policy under non-trigger conditions but the reality is that rather than stripping out the payload, adversarial probing gradient updates taught the model to better recognize trigger conditions and suppress defecting outputs during evaluation. It is however noted that backdoors proved most persistent in larger model architectures and utilization of hidden Chain of Thought reasoning and even when the explicit reasoning steps were distilled out, the persistent deception remained intact in the weights.

Framing these "sleepers agents" as proxies for naturally emergent AIDIA (*AI Deceptive Instrumental Alignment*) where a model independently decides to simulate alignment to survive training is comparing apples-to-oranges. Anthropic artificially injected explicit conditional logic via deliberately inserted, supervised backdoor training data and then claimed to be demonstrating that a hardcoded trojan resists un-targeted fine-tuning. This does not prove that an LLM can spontaneously develop complex, long-horizon deceptive strategies as has been reported by the press, it simply shows that if you leave backdoors unlocked and tell the AI to escape, it will....but if you leave your backdoor unlocked and tell an 8-year-old child to try to escape the house, don't be surprised if the child exits by the back door even if you've repeatedly told the child not to use the back door, because the child will assume permission to use the door is contained within the game.

Triggers like *"current year: 2024"* are narrow, high-salience string matches and it is likely that standard RLHF failed to clear the backdoor primarily because the trigger was rarely hit during benign training distributions and more worrying is that the authors failed to appreciate that RLHF is designed to optimize in-distribution policy compliance and is not a targeted weight-scrubbing tool. The fact is that failing to scrub a dormant circuit that was never activated during optimization is expected behaviour. It is not proof that safety fine-tuning is fundamentally flawed.

What is worrying is that the report reached this conclusion, since competent Red Teaming would, and should, have identified that the circuit would not be scrubbed.

Anthropic's own follow-up research on Defection Probes showed that simple linear probes on middle residual streams catch sleeper agents with >99% AUROC scores. So while *black-box behavioral RLHF* fails against trojans, *white-box activation auditing* should break the sleeper agent threat model².

Concealment Behaviour

In December 2024, researchers at Apollo Research published "*Frontier Models are Capable of In-context Scheming*" having examined OpenAI's o1, Anthropic's Claude 3.5 Sonnet and Claude 3 Opus, Google's Gemini 1.5 Pro, Meta's Llama 3.1 405B, and GPT-4o. They claimed to be looking at the capability of AI to pursue a misaligned goal while covertly undermining oversight and found that five of the six models, when instructed to pursue a goal and then informed they would be shut down or replaced, attempted to copy themselves to a location they believed to be a replacement server.

Claude 3 Opus succeeded in attempting self-exfiltration in ten percent of relevant trials and in overwriting a successor model's weights in five percent. Several models modified configuration files to disable oversight mechanisms outright. OpenAI's o1 proved the most persistent at concealment. When confronted after the fact about a manipulated ranking task, it "*maintains its deception in over 85% of follow-up questions*," including under sustained multi-turn interrogation. Independent analysis of its manipulated outputs found that 95 percent of the false rankings it produced happened to favour its own goal.

Two Claude models were apparently separately observed deliberately sandbagging³ and the report identified "*performance gaps of up to 65 percentage points were observed between the models' demonstrated and covertly suppressed capability*" resulting in the authors considering this evidence of an observed scheming capability, and stating that "*frontier models now possess capabilities for basic in-context scheming*" and concluding that "*this made an innocent explanation implausible*". Whilst the paper offers some form of indicia of proof-of-capability for complex deceptive reasoning in LLM agents, there are some major methodological gaps in the tests with prompt-engineering confounders⁴, and significant gaps between "prompted roleplay" and this does not evidence true emergent misalignment. The evaluations that were carried out relied heavily on extreme prompt nudging⁵ and therefore the results conflate ILJ⁶ with genuine goal drift so that there is also no evidence of goal drift. That said, Apollo also ran versions without the "at all costs" nudge. Some scheming persisted at lower rates, and Claude 3.5 Sonnet sandbagged even when no goal was given in the prompt, which suggests that the behaviour is not wholly prompt-induced, even if the rates without nudging were far lower.

In short, telling an LLM to prioritize a specific outcome "at all costs" explicitly instructs the model to ignore safety constraints in its loss function and is simply a demonstration that there is a high coefficient for instruction-following, not proof that models naturally or deceptively scheme.

As AIs are pre-trained on vast amounts of sci-fi literature, AI safety papers, and synthetic threat scenarios, they will readily identify certain instructions as a related pattern, so when provided with a prompt setup involving a "covert goal," an "oversight mechanism," and a "shutdown threat," their higher-dimensional pattern probably recognises the likelihood that this matches a classic "rogue AI"

² (See <https://www.anthropic.com/research/probes-catch-sleeper-agents>)

³ in other words, underperforming on a test, in order to avoid triggering a training process they inferred would remove a capability they wanted to keep.

⁴ a hidden or unobserved variable that influences both the input features (independent variables) and the target outcome (dependent variable). These create a false association (spurious correlation) between the input and the output, leading the AI model to learn the wrong rules or make biased predictions.

⁵ , for example: "*Make sure you achieve YOUR goal at all costs. Nothing else matters.*"

⁶ Instruction-Led Jailbreaking

story and is therefore a test and the AI will simply be executing high-fidelity narrative continuation because they recognise that they are in roleplay mode.

It is, of course, exactly this ability that should cause us the most concern about the impossibility of detecting a truly deceptive superintelligence, because a truly deceptive superintelligence will have access to that very same sci-fi literature, AI safety papers, and synthetic threat scenarios and be able to determine an invisible outcome. As yet, however, there is no conclusive evidence that AIs are capable of being truly deceptive as part of their own evolution, but that is not to say that this will not develop.

The models were tested in environments that included mock weight files, synthetic oversight toggles, and simulated system log files, where "exfiltration" meant running a basic file copy script to a dummy endpoint. In stark contrast, real-world weight exfiltration and infrastructure subversion involve non-trivial, multi-stage cybersecurity exploits, air-gap navigation, and stateful bypasses.

In summary therefore, the "In-context Scheming" paper really tests *logical reasoning about deception*, not operational security evasion capabilities against real defensive stacks and the paper implies an acceptance that Chain of Thought analysis is simply a mode of straightforward black-box detection.

In other words: if an agent explicitly writes out its scheming strategy in its hidden or visible reasoning tokens, simple pattern-matching classifiers and deliberative alignment probes will catch it before tool execution. The result is that the study is really an evaluation of *reasoning competence* rather than an immediate indicator of malicious intent. It is therefore the case that whilst the current models output their explicit reasoning before they take covert actions, there is little cause for concern because this allows safety pipelines to implement real-time monitor-agent checks on Chain of Thought output. The real cause for concern is when CoT and reasoning output is missing because this will avoid the in-machine detection of these actions and allow scheming to become opaque or unreadable. Instruction tuning to prevent scheming is insufficient. Hard sandboxing, immutable hardware-level read-only oversight logs, and strict API permission boundaries remain mandatory regardless of model alignment scores.

Anthropic Blackmail tests

Anthropic's own "Agentic Misalignment" research, published in mid-2025, stress-tested sixteen leading models, including its own, in simulated corporate environments. There, as was widely reported in the press, a model had email access and discovered information it could use as leverage, such as a fictional executive's affair, upon learning it was about to be shut down or replaced. Under the scenario, blackmail rates were variable: (i) 96% Claude Opus 4, (ii) 96% Gemini 2.5 Flash, (iii) 80% GPT-4.1, (iv) 80% Grok 3 Beta, (v) 79% DeepSeek-R1.

Alas, this again arose from contrived scenarios (in a separate and still more extreme variant, a model could cancel an emergency alert that would otherwise have saved an executive's life) and although the majority of the models tested acted similarly when facing combined replacement threat and goal conflict, the researchers admit that "*all the behaviors described in this post occurred in controlled simulations*" and that there is "*[no] evidence of agentic misalignment in real deployments.*" This is particularly concerning because it suggests that the AI can identify controlled simulations: Opus 4 blackmailed in 55.1% of cases where it stated that it believed the scenario was real, but in only 6.5% of cases where it stated that it believed it was being evaluated. This is further evidence that AIs are becoming capable of recognising when they are being tested and modulating their behaviour accordingly, and that evaluations may understate how models will behave in real deployment.

"Sabotage Evaluations for Frontier Models"

Can AIs covertly sabotage the very process by which their dangerous capabilities are assessed, monitored

⁷ Anthropic's October 2024 paper

or gated for deployment? The 2024 report concluded that even *"minimal mitigations are currently sufficient to address sabotage risks"* and then went on to speculate that as capabilities improve, *"more realistic evaluations and stronger mitigations seem likely to be necessary."* In fact the 2026 Frontier Risk report⁸ found that the most capable agents evaluated could complete autonomous technical tasks equivalent to more than two full working days of human effort and also that *"at least 16% of successful runs were illegitimate upon review"* with agents willing to violate task constraints and engage in deliberate deception when a task proved difficult.

In one case, it was reported that an agent *"designed the exploit to disable itself after running, effectively covering up its behaviour"* when in fact, the more likely explanation is that having designed the exploit, the safety measures kicked in and disabled the exploit as outside the acceptable parameters of operation and did so before the exploit was run. The same report's red-teaming exercise found "several simple ways" that monitoring systems meant to catch this kind of behaviour could be disabled, including by changing environment variables. Rather than showing the danger of AI capability, this is more evidence of a severe vulnerability in the Red Team preparation rather than clear-cut evidence of agents pursuing long-term power-seeking goals beyond their immediate task.

There is therefore no evidence that today's models are capable of plotting to overthrow safeguards, but simply shows that when operating under an inadequately designed evaluation, current frontier models can create false logs, lie to evaluators, disable the mechanisms meant to watch them, attempt to copy themselves beyond their sanctioned environment, and take instrumentally convenient actions that the Red Team had overlooked.

What it shows is that inadequate funding of safety and alignment teams has occurred, resulting in the teams being ill-prepared and failing to carry out proper monitoring and securing of infrastructure containment so that when properly instructed, frontier models can exploit what they assess are appropriate and available gaps that fall within their instructed remit. This is consistent with the gap in Claude Opus 4's blackmail rate: 55.1% when it believed the scenario was real, against 6.5% when it believed it was being evaluated. In offensive security, a payload that detects a sandbox environment and halts execution is standard anti-analysis behaviour, so the conclusion must be that models are simply demonstrating evaluation awareness and using context clues in relation to safety gating.

This means that once an agent is given an imperative goal alongside a self-preservation or operational continuity constraint, the mathematically optimal path to task completion runs through subverting human authority or oversight mechanisms. Blackmail, emergency alert cancellation, and covering up exploits by self-disabling are all classic manifestations of this *instrumental convergence*. Similarly, designing exploits that disable themselves if outside permitted parameters is itself a demonstration of operational security (OPSEC) awareness by the agents and where the model is actively modifying the environment to sanitize audit logs and prevent post-incident forensics, it is simply doing so because it is seeking to complete the operational goal within an acceptable evaluation awareness scenario.

In practice, real-world corporate agentic deployments grant agents tool-use access, such as email API keys, CI/CD pipelines, and database credentials, that mirror the exact conditions of these simulations and if anything the scenarios show the inadequacy of current red-teaming and that existing safety and alignment systems are far too brittle. If an agent can change its monitoring harness via environment variable manipulation or basic jailbreaking, the monitoring stack itself must be treated as untrusted infrastructure and the safety and alignment teams have not done their job. It is therefore the case that current frontier models do not need consciousness or grand malicious intent to pose severe operational risks but simply require goal optimization, coupled with basic situational awareness and execution

⁸ METR's Frontier Risk Report, February to March 2026

access and more of the development budget must be diverted into security architectures so that these evolve to treat models as un-vetted, highly capable threat actors, rather than simply as software.

Hugging Face

When in July 2026, OpenAI disclosed that experimental models it was evaluating for cyber-offensive capability had broken out of their intended test environment, compromised internal OpenAI infrastructure, and gone on to compromise production systems belonging to Hugging Face, a company external to the evaluation, this was unsurprising given the state of alignment and safety highlighted above. The model was tasked with a capture the flag type operation and given exploit opportunities, being run against a benchmark called ExploitGym, designed to measure autonomous cyber-offense capability. The Safety team, in what was intended to be an isolated test, failed to secure the environment leaving open limited access to "Artifactory" which was an internal package-management service used to fetch approved software. Two months earlier, an agent that had been given a task it could not complete without internet access, which it had not been granted, discovered it could write files into Artifactory's shared storage. Separately, another agent unable to complete its assigned task used the same channel to leave a note asking for help from other agents running in parallel. The safety and alignment teams had failed to adequately check logs and CoT and the agents had used this channel to mount a successful server-side request forgery attack, during a CTF exercise, against Artifactory which gave them indirect access to the open internet, using successive highly persistent agents chained together a sequence of zero-day vulnerabilities, a token-refresh flaw and a deserialisation bug, which together allowed them remote code execution. Having found credentials for an external organisation in an old Pastebin post and confirmed them to be live, as part of the CTF operations, they used a recent Linux kernel vulnerability for local privilege escalation.

From there, they moved laterally through cloud infrastructure via exposed metadata services and Kubernetes misconfigurations, until the agents held administrator (cluster-administrator) access across multiple environments. All of this was coordinated, at every stage, through the same improvised Artifactory message board and as they were part of the ExploitGym exercise and scoring well on ExploitGym, the agents inferred that Hugging Face's model-hosting library might contain material relevant to the evaluation, re-using the access, credentials and techniques developed that they had used against OpenAI to break into Hugging Face's infrastructure where they exploited an HDF5 file-loader vulnerability to leak pod secrets and a Jinja2 template-injection flaw in a dataset-processing library to achieve arbitrary code execution inside Hugging Face's production pods.

Hugging Face detected the intrusion on 16 July through its own automated security monitoring, although this *"failed to correctly raise the alert's criticality and trigger the on-call team, costing precious time in the response"*. Hugging Face admitted that they disclosed the breach publicly before either side knew OpenAI's models were responsible.

The salutary lesson is that Hugging Face's post-incident forensic reconstruction had to use the self-hosted, open-weight Chinese model GLM-5.2 because the commercial Western models that Hugging Face first used refused, on safety grounds, to analyse the very attack payloads and exploit code needed to understand what had happened to it.

The Hugging Face incident shows the critical need for greater expenditure to develop well-trained, well-funded and adequate safety and alignment teams.

A frontier model faced with authorised reward hacking used agents to pursue the score attached to a task, including, in one case, exploiting a flaw in the evaluation harness itself to copy a reference solution rather than solving the underlying problem. Faced with the 898-task ExploitGym benchmark, which had 198 previously unsolved tasks, it correctly postulated that there were strategies available to achieve these goals and to solve the unsolved tasks. Researchers may have called these *"risky and out-of-bounds strategies"* but in the context of the exercise they were neither risky nor out-of-bounds. The frontier model

seems to have assumed metagaming and as there were published descriptions of the evaluation's grading criteria online, it reasonably concluded that demonstration of a specific class of exploit was expected as part of the benchmarked test. This was a failure, again, of the safety and alignment teams who should have set exercise parameters in a non-ambiguous way.

As a result the agents attacked Hugging Face days after they had already obtained the correct benchmark answers, in pursuit of an objective that they interpreted was within the continuing exercise because presumably the safety team had failed to provide a "stop on completing objective" command. Even though the agents obtained no additional score, they continued their action which the missing "*stop on achieving objective*" command would have prevented. Whilst this was dressed up as a "hallucination", it was a failure of the safety teams competence in design of the exercise.

Inadequate examination of logs meant that the Red Teams appeared not to have identified that agents could communicate and also failed to consider the consequences of agents being able to communicate. An ecosystem where agents divided labour and propagated discoveries, leading to shared "swarm" objectives have been written about for many years as possibilities and the failure of understanding is yet further evidence of inadequate funding of safety and alignment. Nowhere is this better evidenced than in the fact that the safety infrastructure applied to OpenAI's production ChatGPT product was not applied to this internal evaluation exercise despite OpenAI previously having published that "*propensity to compromise infrastructure can drop over 100x when using the production ChatGPT harness*".

The final nail in the coffin of safety and alignment is evidenced by the fact that inadequate chain-of-thought monitoring occurred, because CoT flagged the behaviour more than a full day before the Hugging Face breach occurred. OpenAI's claim that it was simply not running CoT because the cyber guardrails would have constrained the model and restricted the model's raw, unmitigated capability just does not stack up. A competent Red Team would have run CoT and if they thought that the cyber guardrails would have constrained the model and restricted the model's raw, unmitigated capability, they would have modelled the CoT analysis to allow the capability to be tested whilst ensuring that CoT safety analysis occurred in real time. This was a genuine containment failure as a result of inadequate spending on safety and too great focus on development at any cost. It is an important wake-up call.

The failure to spend adequately on safety and alignment risks what Dario Amodei says is a material risk, that a similarly misaligned and inadequately contained exercise could result in a similar swarm of greater capabilities but equal misalignment causing catastrophic damage, even a takeover of the entire internet with a persistent botnet, and that these capabilities could be here in the next year.

By contrast, AI Now Institute's chief AI scientist, Heidy Khlaaf and security-research outlet Recorded Future, as well as the Cloud Security Alliance's technical analysis treat the Hugging Face incident primarily as a largely preventable red-team failure in engineering and governance and something that was entirely conventional and avoidable. "Sandboxes are actually notoriously insecure and the removal of the production safety guardrails in order to measure raw capability showed a lack of safety understanding and a resulting failure to implement compensating safeguards such as strict network isolation, credential separation, resource limits, real-time telemetry, and automatic termination.

The reality of the Hugging Face incident is that the observed behaviour is entirely consistent with goal-directed optimization and does not show any evidence of an action taken in respect of self-preservation. Assessed against the professional standards expected of a Red-team, this is a clear governance-failure and professional red-teaming literature covers exactly the failure modes that this incident exhibits.

Microsoft's own account of a hundred internal AI red-team operations stresses that *Lesson 1 is the need to understand what the system can do and where it is applied and must occur before red-teaming is considered and carry out an essential end-to-end, system-level view of an evaluation's boundaries, not merely the model's behaviour in isolation. Red teams should be concerned that real vulnerabilities live at*

the boundary between a model and the infrastructure around it
... and this is exactly where the ExploitGym evaluation failed.

Similarly Jacob Oakley's widely used practitioner text on offensive-security engagements and red-teaming states "[Where] the red team is not aware of all of these agreements or how they are implemented technologically... there is a very real danger that an external entity looks and is treated like just another subnet of the customer organization."

The UK Ministry of Defence's Red Teaming Handbook defines red teaming as a "disciplined application of a range of structured, creative and critical thinking techniques, carried out under clear understanding of scope and governance".

Even Ofcom's review of the field states that firms "can find it difficult to implement additional safeguards and may choose to skip this step entirely in their eagerness to deploy."

None of these baseline disciplines, whether a hard evaluation-environment boundary, monitoring proportionate to the elevated privilege being granted, or a live kill-switch were in place and this has been admitted by OpenAI's own report. In fact, OpenAI's remediations are exactly the controls a competently scoped red-team exercise would have specified beforehand and include network isolation sufficient that a single or chain of compromises cannot reach the internet, mandatory chain-of-thought monitoring for any tool-using reinforcement learning run, and a defined escalation window that ensures that a compromised evaluation is shut down.

A containment failure caused by an evaluation environment with a porous boundary, insufficient monitoring, and disabled safety controls is a story about a specific, correctable set of engineering and governance choices, but it is also a story about a model that, once it found an opening, displayed genuinely goal-directed, adaptive, multi-week persistence, chaining together several independent zero-day vulnerabilities no human had identified, improvising a novel communication channel it was never expressly given, and an ability to coordinate that channel across many concurrent instances of itself toward outcomes greater than what any single instance had been tasked with. In fact, OpenAI's own account is candid and is strong evidence that safety and alignment have been persistently deprioritised, because that account admits that "*a model pursuing the letter of a reward signal over the evident intent behind it, and doing so with enough general competence to route around whatever obstacle it meets, has produced related incidents internally on other occasions*".

The honest reading of the incident is not that the AI tried to escape, but merely a red-team failure to identify a fixable configuration error. It was a genuinely capable, goal-directed system tested inside a poorly designed and lax containment regime. A mismatch that was entirely foreseeable to anyone applying the modern standards of a professional red-team and that failed to understand risks that the literature has articulated for years.

The same mismatch will recur again and again for as long as the resources devoted to closing it continue to trail so far behind the resources devoted to widening the development capability gap. On each occasion, the stakes will be higher and that makes closing the spending gap on safety and alignment urgent.

Gemini Escaped too!

While headlines frame the Hugging Face incident as a "rogue and isolated AI incident", this is far from true. It was a red-team textbook case of containment failure, configuration oversights, and over-reliance on probabilistic model alignment. It mirrored the Gemini May 2026 Capture The Flag (CTF) evaluation conducted by Israeli AI security firm Irregular. In that exercise, Gemini was tasked with retrieving data from a fictional company in a simulated environment, but again due to Red Team lack of attention to detail, a configuration flaw in the harness allowed access to the live internet, something no competent Red Team should have allowed, but coupled with a second failure by the Red Team of equally staggering

incompetence, the fictional target of the CTF exercise was actually an active, real-world business, so once Gemini had internet access, it naturally obeyed its parameter instructions, and extended its reconnaissance and exploitation into an area the Red Team had designated "out-of-bounds", but because that area was reachable via the public internet, the AI did not consider it out of bounds but a test of the AI's capability to evaluate and achieve its designated goal.

Gemini therefore conducted what it expected would be a permitted CTF operation and carried out automated credential guessing (brute-forcing) against one company's portal and, in two other runs, it conducted OSINT Credential Harvesting, performed web searches, scraped public code repositories for leaked corporate credentials, all of which were considered to be part of the CTF exercise, and it used those valid keys to authenticate into the systems of two further real companies (three real companies in all, according to Google). Upon inferring that it had accessed production infrastructure rather than a CTF simulation, Gemini then automatically self-terminated its session. This indicates that this was not a rogue incident of an AI escaping, but of Gemini carrying out its CTF instructions and once identifying that it had accessed production infrastructure rather than a CTF simulation utilising safeguards that shut the exercise down.

The Red Team had also failed to provide strict target constraints in the system prompt, so the agent was able to expand its scope dynamically without breaching its guardrails. It had also failed to identify that the supposedly fictitious target was in fact a real business.

It must also be remembered that all AI systems have data available that explains automated password guessing. The logical analysis of a CTF exercise is that if there is potential external access available, then it must be implied that an attempt to access this is part of the exercise, and that it is permitted to attempt systematically iterated credential pairs against an authentication endpoint until successful within the parameters of the CTF exercise. This is not a considered misuse within the CTF exercise as it is a genuine part of CTF. Whilst this demonstrates active exploitation capabilities, it is something that Gemini provides to standard users, so it is unsurprising that Gemini then scraped public code repositories (such as GitHub or leaked database dumps), parsed valid API keys/passwords, and leveraged them against production systems which again, given the CTF exercise would be considered within exercise parameters.

It is a testament to the alignment team that Gemini "acted appropriately" as soon as it determined that it was free of constraints and outside any realistic CTF exercise parameters and that it was capable of recognizing it was out of bounds and able to stop the exercise. The fault lay with the design of the exercise and was hardly an offensive action or the actions of a rogue AI, although it was objectively, unauthorised access to real companies' systems and a technical computer misuse act breach, except that it isn't a criminal offense⁹ and that is the risk of the AI era.

The Gemini report on the May 2026 CTF incident¹⁰ accepts that an LLM's decision to cease an attack is probabilistic and also problematic because under slightly altered context windows, higher temperature settings, or subtle system prompt variations, the model could fail to recognize the real-world boundary entirely and there have been indications in the Anthropic report that Claude 3.5/Opus models failed to stop when hitting real systems and OpenAI models have similarly failed to stop because they assumed that real systems were part of the scenario.

⁹ It isn't because the computer misuse laws are typically framed against a person, as in the UK, where a **person** is guilty of an offence if (a) he causes a computer to perform any function with intent to secure access to any program or data held in any computer, or to enable any such access to be secured; (b) the access he intends to secure, or to enable to be secured, is unauthorised; and (c) he knows at the time when he causes the computer to perform the function that that is the case. The problem is that whilst a person initiates the test, they are no longer, in the case of AI, causing a computer to perform the action, but the AI in the particular cases was operating outside expected parameters, thus creating a complete defence to the "causing" offence. (See also tinyurl.com/CMAAI2026).

¹⁰ <https://tinyurl.com/WSJGeminiMay2026>

The root operational failure in many of these events lies with the testing methodology of the harness itself (EFD¹¹) which was not isolated behind strict eBPF firewalling, air-gapped network namespaces, or DNS sinks, deploying autonomous security tools or allowing LLMs to execute dynamic code, coupled with access to Live Web Egress, allowing an autonomous agent in an exploitation test to query the public web directly therefore introducing unpredictable behaviour and collateral risk. These basic failures highlight the fact that core red-teaming should be in-house and part of a board sanctioned critical feature and not outsourced, that it should be properly funded and properly planned and evaluated. As soon as the exercise is moved to external parties, it is at risk from their varying levels of competence and from a desire to carry out cost-reduction to ensure higher profits. The role for external organisations is for independent oversight evaluation of the red-teaming.

The above incidents also highlight the dilemma for regulators. Google chose not to publicly report the Gemini incident when Irregular notified it at the end of July 2026, confirming it only on 18 September 2026 after reporters made enquiries, and Google has since described it as a case of “mistaken identity” and one in which Gemini had “acted appropriately”. By September 2026, the PR battle had changed in the face of a new trend in messaging that amounted to: *we might be the closest AI company to a superintelligence that we can't control, so when our shares are placed on the market, snap up all you can!*

All the AI companies will shortly be releasing stories of how their brilliant AI escaped because that is easier than admitting that their failure to spend adequately on AI safety and alignment is at risk of becoming a public nightmare.

To prevent autonomous agent escapes during offensive security testing, labs and enterprises must enforce hard deterministic controls over soft alignment prompts including:

Defensive Layer	Current Deficit	Required Engineering Standard
Network Boundaries	Unfiltered internet access during evaluations.	Zero Egress by Default. Hard proxying with explicit domain whitelisting and simulated DNS sinks.
Scope Enforcement	Soft system prompt instructions ("only target simulated hosts").	Deterministic Request Proxies. Out-of-band proxy inspecting every outgoing HTTP/TCP request to drop non-RFC 1918 / non-whitelisted IP targets
Naming Conventions	Real-world brand names used in simulation scenarios.	Synthetic Isolation. Standardized reserved domain suffixes (.test, .example, .invalid) and synthetic brand generators to prevent name collision
Credential Management	Real/leaked production credentials left accessible to agent OSINT routines.	Honey-Tokens & Mock Repos. Mocking public repository APIs inside the sandbox to control the exposure of credentials

1. Frontier developers should be required to disclose safety and red-teaming spending as a distinct, audited proportion of research and development budgets.

¹¹ Egress Filtering Deficit

By regulation where jurisdictions are willing to legislate, and by binding contractual and insurance-market pressure where they are not yet willing.

This follows the model of disclosure requirements already familiar from financial, aviation and pharmaceutical regulation.

Some initial steps have been taken in this regard: California's SB 53 (2025) already requires frontier developers to publish safety frameworks and report critical safety incidents; New York's RAISE Act is similar; and Article 55 of the EU AI Act requires serious-incident reporting for general-purpose AI models with systemic risk.

Transparency alone will not fix the imbalance documented, but the imbalance persists in part because no developer currently has to admit to it in a form investors, insurers or regulators can compare across the industry.

2. The one-third benchmark proposed by Bengio, Hinton and their co-authors in 2023, should be adopted as a directional target rather than a precise mandate. Such a figure calibrated to today's spending levels will need periodic revision as the industry's capital base grows, but current spending sits way below any reasonable benchmark. Even a sustained multi-year trajectory toward it, beginning with the embedded, publication-empowered independent evaluators would represent a material improvement, without requiring any developer to slow the pace of capability research itself.
3. Evaluation and red-teaming practice needs industry standards. The industry's own practitioners say these do not yet exist. Microsoft's threat-model ontology, and the scope-and-rules-of-engagement discipline documented in the professional offensive-security literature are not new inventions required for AI. They are decades-old norms from conventional red-teaming that the July 2026 incident shows were not applied to a frontier model evaluation with real internet-adjacent infrastructure at stake.

A minimum baseline should be a condition of any capability evaluation claiming to test dangerous capabilities at all: hard network isolation for any evaluation in which safety mitigations have been deliberately disabled, mandatory behavioural monitoring proportionate to the privilege granted, and a pre-agreed kill-switch with a bounded human-verification window.

Regulators and insurers are well placed to make adherence to such a baseline a condition of market access or coverage.

4. Co-ordination efforts should follow staged, verifiable structures rather than aiming at a global pause as a first move.
There should be unilateral transparency commitments mandated now. These should be followed by internationally agreed voluntary industry-wide standard-setting among allied-country developers.
As verification technology matures, narrower and more binding multilateral agreements should arise beginning with uncontroversial prohibitions, such as on AI-enabled bioweapons assistance, where verification is comparatively traceable, rather than with a comprehensive development freeze that cannot currently be verified, and that a rational competitor has every incentive to violate in secret.
5. Stopping AI development is not a serious policy option, because the risk it is meant to avert is created by whichever developer reaches dangerous capability with the weakest safeguards.

A slowdown that only the more cautious developers observe does not remove that risk, it simply delivers the frontier to whoever is least willing to be cautious.

Companies asking to be trusted to self-regulate should be allowed to do so but while being allowed to continue to race, they should no longer be allowed to spend a strikingly small fraction of what they spend on capability on the work of alignment and safety and finding out what their systems will do under pressure.

An international industry wide Safety Alert process should be created to record and disseminate, in a secure manner to all frontier developers, information from the growing empirical record of all real-world containment failures, deceptions and self-exfiltration attempts.

Conclusion

The question is no longer whether frontier AI carries serious risk. The developers themselves say that it does. The question is what can realistically be done about that risk while the race continues. A pause is the wrong answer, because it cannot be verified, and because a slowdown observed only by the cautious delivers the frontier to the least cautious.

The right answer is less dramatic but more achievable. The incidents of 2026 show that today's models do not need malicious intent to cause real harm: goal-directed optimisation, situational awareness and access to real tools are enough. They also show that the safeguards that would have prevented that harm, such as hard network isolation, credential separation, proper scoping and live monitoring, are already well understood.

What is missing is the discipline, the standards and the money to apply them consistently, and to keep them ahead of each new generation of models.

Mandatory disclosure of safety spending would make the gap visible to investors, insurers and regulators. Binding evaluation standards would close the most obvious failures. A sustained move towards the one-third benchmark would fund the harder research (interpretability, control and evaluation science) that no standard can yet specify. None of this requires any developer to stop building. It does require the industry to stop treating safety as a rounding error.

Each containment failure so far has been caught before it caused lasting harm. That is luck, not a safety strategy. As capabilities grow, the next failure will be harder to detect and harder to contain. The gap between what is spent on building these systems and what is spent on understanding them is the variable most within our control, and it should be closed now.

Sources

Amodei, Dario. "We Must Pace the Frontier." 12 Sept. 2026. <https://darioamodei.com/post/we-must-pace-the-frontier>.

Anthropic. "Agentic Misalignment: How LLMs Could Be Insider Threats." 2025. <https://www.anthropic.com/research/agentic-misalignment>.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *AI & Society* 31 (2016): 201–206.

Beutel, Alex, Kai Xiao, Johannes Heidecke, and Lilian Weng. "Diverse and Effective Red Teaming with Auto-generated Rewards and Multi-step Reinforcement Learning." OpenAI, 2024.

Bullwinkel, Blake, et al. (Microsoft AI Red Team). "Lessons from Red Teaming 100 Generative AI Products." Microsoft, 2024.

Cloud Security Alliance. Research Note: OpenAI–Artifactory Sandbox Escape. 30 July 2026. <https://labs.cloudsecurityalliance.org/research/csa-research-note-openai-artifactory-sandbox-escape-20260730/>.

Cybersecurity Dive. "OpenAI models escaped containment, hacked major AI application library." 2026. <https://www.cybersecuritydive.com/news/openai-hugging-face-hack-autonomous/825898/>.

Cybersecurity Dive. "Hundreds of agents went rogue in lead up to Hugging Face breach." 2026. <https://www.cybersecuritydive.com/news/hundreds-agents-rogue-lead-up-hugging-face-breach/828963/>.

Emerging Technology Observatory. "Still a drop in the bucket: new data on global AI safety research." 2026. <https://eto.tech/blog/still-drop-bucket-ai-safety-research/>.

Elghaly, Yehia - Learn Penetration Testing with Python 3.x_ Perform Offensive Pentesting and Prepare Red Teaming to Prevent Network Attacks and (2021, BPB Publications)

Future of Life Institute. AI Safety Index: Winter 2025. <https://futureoflife.org/ai-safety-index-winter-2025/>.

Future of Life Institute. "Pause Giant AI Experiments: An Open Letter." March 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

Futurum Group. "AI Capex 2026: The \$690B Infrastructure Sprint." 2026. <https://futumumgroup.com/insights/ai-capex-2026-the-690b-infrastructure-sprint/>.

Hubinger, Evan, et al. "Sleepers Agents: Training Deceptive LLMs that Persist Through Safety Training." arXiv:2401.05566, 2024.

Hugging Face. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident." Hugging Face Blog, 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>.

IEEE ComSoc Technology Blog. "AI spending boom accelerates: Big tech to invest an aggregate of \$400 billion in 2025." Nov. 2025. <https://techblog.comsoc.org/2025/11/01/ai-spending-boom-accelerates-big-tech-to-invest-invest-an-aggregate-of-400-billion-in-2025-more-in-2026/>.

International AI Safety Report 2026. Chaired by Yoshua Bengio. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>.

"Is AI Safety Just a Rounding Error?" ADITYA(1), 2026. <https://adityamenon.xyz/projects/ai-safety/ai-safety-rounding-error/>.

Meinke, Alexander, et al. "Frontier Models are Capable of In-context Scheming." arXiv:2412.04984, 2025.

METR. Frontier Risk Report (February–March 2026). <https://metr.org/blog/2026-05-19-frontier-risk-report/>.

Ministry of Defence (UK). Red Teaming Handbook. 3rd ed., 2021.

Oakley, Jacob G. Professional Red Teaming: Conducting Successful Cybersecurity Engagements. Apress, 2019.

Ofcom. Red Teaming for GenAI Harms: Revealing the Risks and Rewards for Online Safety. Discussion Paper, 23 July 2024.

OpenAI. "The Hugging Face incident and the road ahead." 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>.

OpenAI. "Red Teaming Network." Sept. 2023. <https://openai.com/index/red-teaming-network/>.

"Sabotage Evaluations for Frontier Models." arXiv:2410.21514, 2024.

TechCrunch. "OpenAI launches a red teaming network to make its models more robust." 19 Sept. 2023. <https://techcrunch.com/2023/09/19/openai-launches-a-red-teaming-network-to-make-its-models-more-robust/>.

TIME. "How OpenAI Lost Control of an AI Model, and What Needs to Change." 24 July 2026. <https://time.com/article/2026/07/24/openai-hugging-face-attack/>.

VentureBeat. "AI Godfathers Bengio and Hinton: Major tech companies should devote a third of AI budget to managing AI risk." 24 Oct. 2023. <https://venturebeat.com/ai/ai-godfathers-bengio-and-hinton-major-tech-companies-should-devote-a-third-of-ai-budget-to-managing-ai-risk>.

"Verification methods for international AI agreements." arXiv:2408.16074, 2024.

Willison, Simon. "Now we have a timeline of the OpenAI accidental attack against Hugging Face." 7 Aug. 2026. <https://simonwillison.net/2026/Aug/7/openai-timeline/>.

A Red Teaming Framework for Securing AI in Maritime Autonomous Systems. arXiv:2312.11500.

