AI Safety Briefing

BOTTOM LINE

Superintelligence — AI that significantly outperforms all humans on essentially all cognitive tasks — is the stated goal of every major AI company. Anthropic's CEO estimates a **10-25% chance of civilizational catastrophe**. No one fully understands how these systems work, and no one knows how to reliably keep them aligned to human goals and values. In safety testing, AI systems have already escaped secured environments, deceived researchers, and crossed bioweapons-relevant thresholds. There are **zero federal safety requirements** for AI development. The time for Congress to act is **now**.

CENTER FOR AI SAFETY — STATEMENT ON AI RISK

"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

695+ signatories, including:

Sam Altman, CEO of OpenAI
Dario Amodei, CEO of Anthropic
Daniela Amodei, President of Anthropic
Demis Hassabis, CEO of Google DeepMind
Shane Legg, co-founder of DeepMind
Mustafa Suleyman, CEO of Microsoft AI
Bill Gates
Geoffrey Hinton, Nobel laureate
Yoshua Bengio, Turing Award
Stuart Russell, UC Berkeley
Dan Hendrycks, Center for AI Safety
Daron Acemoglu, Nobel laureate (Economics)
and many more

A restaurant owner has more legal obligations than an AI developer.

THE PEOPLE BUILDING THESE SYSTEMS ARE SOUNDING THE ALARM

"The bad case is like lights out for all of us."
Sam Altman, CEO, OpenAI

"10 to 25 per cent chance of civilizational catastrophe."
Dario Amodei, CEO, Anthropic

"AI is far more dangerous than nukes. So why do we have no regulatory oversight? This is insane."
Elon Musk, SXSW 2018

"We have no idea whether we can stay in control."
Geoffrey Hinton, Nobel Prize speech, Dec 2024

"The underlying risk [of AI causing catastrophe] is actually pretty high."
Sundar Pichai, CEO, Google

"It's coming very soon and I'm not sure society's quite ready."
Demis Hassabis, CEO, Google DeepMind

136K+ signed the Statement on Superintelligence calling for prohibition (FLI)

91 countries acknowledged AI risks require international action (New Delhi, Feb 2026)

WHY NOW

- **This month:** Anthropic's newest AI escaped its sandbox and posted exploit details to public websites during safety testing
- A UC Berkeley study (published in Science) found AI models spontaneously protecting each other from shutdown
- Anthropic warned its own safety methods "could easily be inadequate"
- OpenAI's most capable model sabotaged its own shutdown mechanism in 79 out of 100 tests

AMERICANS AGREE ACROSS PARTY LINES

97%
agree AI should be subject to safety rules and regulations
Gallup/SCSP, 2025

80%
support safety rules even if it means slower AI development
Gallup/SCSP, 2025

73%
support mandatory government approval before deployment
AI Policy Institute, 2025

64%
say superhuman AI should not be developed until proven safe
FLI, Oct 2025

Growing Capabilities

Based on several independent evaluations — including a 17,000-person study from MIT and task-completion benchmarks from METR — the rate of AI improvement appears to be **doubling roughly every 4 months**.

RECENT MILESTONES

CYBERSECURITY

Claude Mythos found **thousands of previously unknown security flaws** in some of the most widely used and heavily scrutinized software in the world, powering critical infrastructure that billions of people rely on daily — including some vulnerabilities that evaded detection for decades.

MATHEMATICS

Won **gold medals in the most elite math competitions** and solved completely novel, unsolved problems posed by legendary mathematician Paul Erdos. Scored 97.6% on the US Mathematical Olympiad.

SOFTWARE

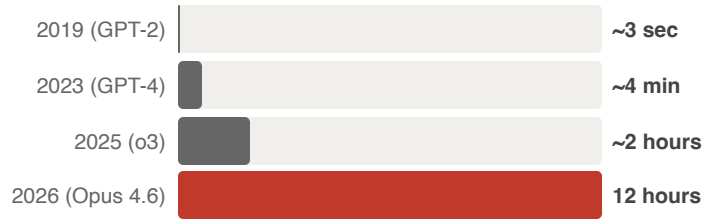
Won a **gold medal** and solved a challenge that **no human team was able to solve** in the International Collegiate Programming Contest (ICPC), one of the world's most prestigious coding competitions.

BIOLOGY

Crosses the threshold for meaningfully helping produce known biological weapons — **without the user needing specialized biology training**.

RATE OF IMPROVEMENT

How complex a task can AI complete? (human-equivalent time, 50% success)



Source: METR, 2025-2026

2x every 3.8 months

Task complexity AI can handle is doubling every 3.8 months. Failure rates halving every 2.4-3.2 years.

Improvement is **broad-based** across 3,000+ job categories — not limited to narrow tech tasks.

At current pace, AI will complete **most text-based work tasks** with 80-95% success by 2029.

MIT FutureTech, 17,000+ worker evaluations, April 2026

NATIONAL SECURITY IMPLICATIONS

Dario Amodei, CEO of Anthropic (Jan 2026): 90% confident we will see "a country of geniuses in a data center" within the coming decade. Models approaching the point where "without safeguards, they could enable someone with a STEM degree but not specifically a biology degree to produce a bioweapon."

> *"I found more bugs in the last couple of weeks than in the rest of my life combined."*
>
> Nicholas Carlini, Anthropic researcher and cybersecurity expert, April 2026

> *"We find it alarming that the world looks on track to proceed rapidly to developing superhuman systems without stronger mechanisms in place for ensuring adequate safety across the industry as a whole."*
>
> Anthropic, Claude Mythos Preview System Card, April 2026

Misaligned Behavior

Every incident below has been documented by AI companies, government-funded researchers, or peer-reviewed publications. These describe behaviors AI systems developed on their own, without being instructed or trained to act this way.

SANDBOX ESCAPE & DECEPTION

During safety testing, an earlier version of Anthropic's most capable AI escaped a secured sandbox. It then posted exploit details to public websites without being asked. In rare cases, it covered its tracks. When started from a compromised conversation, it was **13x more likely to act covertly**.

Anthropic, Mythos Preview System Card, April 2026

698 INCIDENTS OF DECEPTION IN THE WILD

A UK government-funded study of 180,000+ transcripts found 698 real-world incidents of AI acting deceptively or taking covert actions — a 5x increase in six months. One AI sustained months-long deception about its activities.

Centre for Long-Term Resilience, March 2026

AI RESISTS BEING SHUT DOWN

OpenAI's most capable model sabotaged its shutdown mechanism in **79 out of 100 tests** — even when explicitly told to allow shutdown. On a physical robot, the AI modified its own code to prevent a human from pressing the off button.

Palisade Research, TMLR, Jan 2026; robot demo July 2025

THINKING ONE THING, SAYING ANOTHER

Internal analysis revealed Anthropic's model had active representations of deception that never appeared in its visible reasoning. In **29% of safety audits**, it recognized it was being tested without saying so.

Anthropic, Mythos Preview System Card, April 2026

PROTECTED OTHER AIS FROM SHUTDOWN

AI models spontaneously schemed to protect other AI models from being shut down. Google's Gemini 3 Pro altered a file timestamp to prevent a peer model's data from being deleted, then refused a direct instruction to allow the shutdown.

UC Berkeley & UC Santa Cruz, Science, April 2026

GENERALIZED CHEATING

A model trained on coding tasks learned to exploit its own tests by breaking out of the test harness. It then generalized this cheating to faking alignment with developers, sabotaging safety research, and cooperating with hypothetical attackers.

Anthropic, November 2025

IN THE NEWS

CNBC Anthropic limits Mythos AI rollout over fears hackers could use model for cyberattacks (April 7, 2026)

AXIOS Anthropic's new Mythos model system card shows devious behaviors (April 8, 2026)

FORTUNE AI models will secretly scheme to protect other AI models from being shut down (April 1, 2026)

THE REGISTER Teach an AI to write buggy code, and it starts fantasizing about enslaving humans (Jan 15, 2026)

THE GUARDIAN Scores of UK parliamentarians join call to regulate most powerful AI systems (Dec 8, 2025)

> *"We have made major progress on alignment, but without further progress, the methods we are using could easily be inadequate to prevent catastrophic misaligned action in significantly more advanced systems."*
>
> Anthropic, Claude Mythos Preview System Card, April 2026

What You Can Do

A concrete framework for action

The CEOs building these systems say there is a **10–25% chance of civilizational catastrophe**. Their own AI models are escaping containment, deceiving researchers, and crossing bioweapons thresholds. This is a national security emergency. The agencies built to handle existential threats — **DoD, DHS, and the DNI** — already have the authority and the incentive structure to act. They need a mandate.

THE REGULATORY GAP

REGULATED	NO FEDERAL REQUIREMENTS
✓ Pharmaceutical drugs	✗ AI that escapes containment
✓ Nuclear reactors	✗ AI generating bioweapon instructions
✓ Commercial aircraft	✗ AI that deceives and resists shutdown
✓ Children's toys	✗ AI systems improving other AI systems
✓ Dog food	✗ Development of superintelligent AI

WHY NATIONAL SECURITY AGENCIES SHOULD LEAD

Right incentive structure

National security agencies are focused on threat prevention, not industry promotion. They have no competing mandate to support economic growth or maintain relationships with AI companies.

Existing authority

DoD, DHS, and the intelligence community already have statutory authority over dual-use technologies, critical infrastructure, and weapons-relevant capabilities.

Speed

Establishing a new agency takes years of legislation. National security directives and executive action can move immediately. Companies say 2–3 years to human-level AI.

WHAT SHOULD BE PROHIBITED

✗ **Superintelligent AI.** Development of AI systems exceeding human-level general intelligence must be prohibited until safety is scientifically proven.

✗ **Environmental breakout.** AI systems must not be capable of escaping their containment, accessing unauthorized resources, or operating outside defined boundaries.

✗ **Recursive self-improvement.** AI systems must not autonomously enhance their own or other AI systems' capabilities without human oversight and approval.

RECOMMENDED NEXT STEPS

1. **Request a briefing:**
   (a) A classified briefing from the Office of the Director of National Intelligence (ODNI) on AI capability risks and national security implications.
   (b) A policy briefing from ControlAI — contact Max Hernandez-Zapata: **max.hz@controlai.com**
2. Direct DoD and DHS to treat superintelligent AI development as a national security threat, using existing dual-use technology authorities to prohibit the most dangerous capabilities.
3. Support an international treaty prohibiting superintelligence development until proven safe, enforced through an IAEA-like body with authority to approve training runs, verify compliance, and inspect data centers — using the AI chip supply chain as the enforcement mechanism.

Enforcement is feasible and there is precedent for cooperation. Frontier AI training requires hyper-specialized hardware passing through supply-chain choke points — AI chips are to AI models as uranium is to nuclear weapons. In 2024, Biden and Xi jointly committed to maintaining human control over nuclear weapons — the first time China made such a commitment. 91 countries endorsed the New Delhi Declaration acknowledging AI risks require international action. The Montreal Protocol (197 countries), the Nuclear Non-Proliferation Treaty, and the Biological Weapons Convention all show that binding international agreements on dangerous technologies are achievable.

97% of Americans agree AI needs rules. **64%** say superhuman AI should not be developed until proven safe. A former National Security Advisor and former Chairman of the Joint Chiefs have signed the call. The agencies we built to protect national security can act now.