AI is one of the most important things humanity is working on. It is more profound than, I dunno, electricity or fire.

Sundar Pichai, CEO of Google, 2018

What is Safety? Corporate Discourse, Power, and the Politics of Generative AI Safety

Ankolika De, Pennsylvania State University
Gabriel Lima, Max Planck Institute for Security and Privacy
Yixin Zou, Max Planck Institute for Security and Privacy

CHI'26
Barcelona, Spain | April 13 - 17, 2026

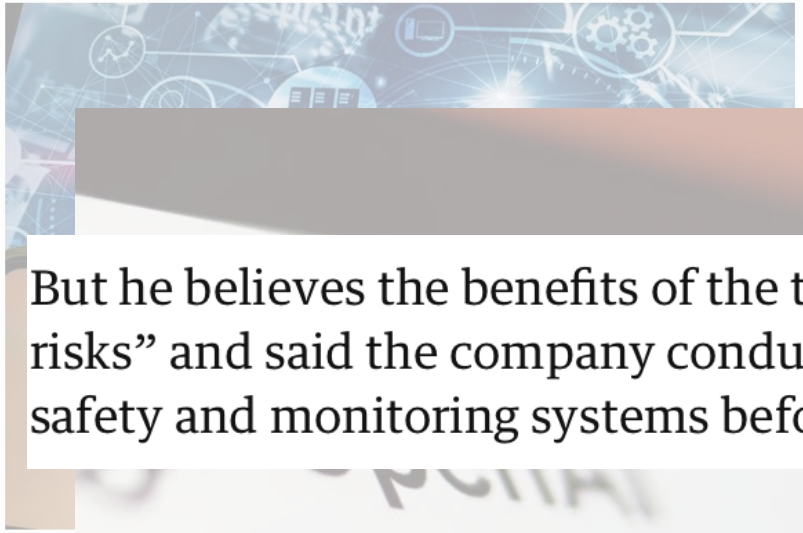
MAX PLANCK INSTITUTE
FOR SECURITY AND PRIVACY

PennState
College of Information
Sciences and Technology

AI Safety

Anthropic's AI sparks concerns over a new national security risk

By AARON MAK | 04/09/2026 04:53 PM EDT



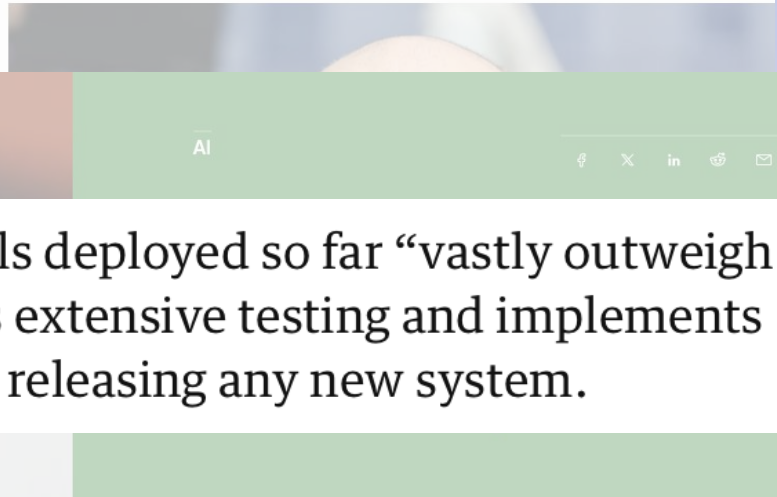
Anthropi
Images

AI and
large lan

On Tuesday, Anthropic announced Claude Mythos Preview, a model the company claims is capable of exploiting vulnerabilities in every major operating system and internet browser. According to Anthropic, more than 99 percent of the thousands of vulnerabilities that Mythos has identified aren't patched yet, and many have gone unnoticed for decades. Anthropic is now launching Project Glasswing, an initiative that involves collaborating with companies like AWS and Google to safely deploy these AI capabilities and enhance cybersecurity.

Google told staff worried about Pentagon AI deals that the company is 'leaning more' into national security contracts

By Hugh Langley +



AI

Lauren Forristal — 8:23 AM PDT · April 8, 2026

- In an internal all-hands, Google DeepMind leaders addressed staff concerns about Pentagon work.
- Leaders said there was a "robust process" to ensure the contracts align with Google's AI principles.
- At the same time, leaders said Google was pursuing more contracts in areas like cybersecurity and biosecurity.

But he believes the benefits of the tools deployed so far "vastly outweigh the risks" and said the company conducts extensive testing and implements safety and monitoring systems before releasing any new system.

Background

Gen AI is shaped by **social actors, institutions, and existing power structures** (Strohmayer et al 2022).

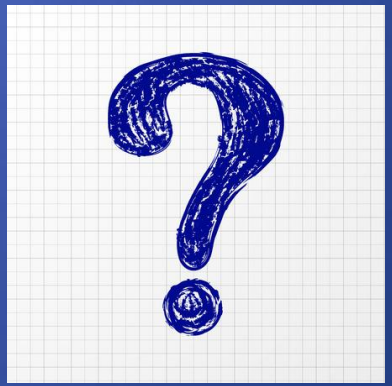
Technical interventions alone are **insufficient** to address AI (Aven & Ylonen, 2018).

“Influential” discourses on safety can shape **public perception, policy, and research priorities** (Coldicutt, 2024).

Repeated narratives stabilize knowledge and produce legitimacy. (Jager & Maier, 2014)

AI Safety, where safety goes beyond the absence of harm or risk, but also as the ***intertwining of inclusion, equitable access, transparency, and accountability*** (Stardust et al., 2023).

How do GenAI companies construct the notion of safety in their public communications?



Methods

N= 25/platform

Blogs, Press releases, Policy statements

Critical discourse analyses (Jäger & Maier, 2016)

Discourses

Dominant discourses reflect existing power structures and influence whose perspectives are valued (van Dijk, 2017)

How we think about safety and alignment

The mission of OpenAI is to ensure artificial general intelligence (AGI) benefits all of humanity. Safety—the practice of enabling AI's positive impacts by mitigating the negative ones—is thus core to our mission.

Our understanding of how to advance safety has evolved a lot over time, and this post is a current snapshot of the principles that guide our thinking. We are not certain everything we believe is correct. We do know AI will transform most aspects of our world, and so we should think through the benefits, changes, and risks of this technology early.

Findings

1. *What* Do Companies Talk About When They Talk About Safety?

- Responsibility and Accountability
- Governance, Oversight and Control
- Risk and Mitigation

2. *How* Do Companies Talk About AI Safety?

- With Authority
- As a Dynamic and Global Construct
- Through Metaphors

1. ***What*** Do Companies Talk About When They Talk About Safety?

"Whatever ***regulations*** we arrive at should be as surgical as possible" (**Anthropic**)

"Any industry where there are potential ***harms*** needs ***evaluations***" (**Anthropic**)

"…developing and delivering ***boldly and responsibly*** …have the potential to assist and ***improve lives of people everywhere*** — this is what compels us" (**Google**)

Findings

1. *What* Do Companies Talk About When They Talk About Safety?

- Responsibility and Accountability
- Governance, Oversight and Control
- Risk and Mitigation

2. *How* Do Companies Talk About AI Safety?

- With Authority
- As a Dynamic and Global Construct
- Through Metaphors

2. ***How*** Do Companies Talk About AI Safety?

Authoritatively

"To be effective at addressing AGI's impact on society, **OpenAI must be on the cutting edge of AI capabilities**--- policy and safety advocacy alone would be insufficient." (**OpenAI**)

"Retired U.S. Army General Paul M. Nakasone has joined our Board of Directors…Nakasone's appointment **reflects OpenAI's commitment to safety and security**.." (**OpenAI**)

2. *How* Do Companies Talk About AI Safety?

Safety is Never Finished…

…we cannot predict all the ways people will use our technology, **nor all the ways people will abuse it**… we believe that **learning from real-world use is a critical component** of creating … increasingly safe AI. (OpenAI)

… Harnessing the potential of AI is grounded in our founding mission to **organize the world's information and make it universally accessible and useful…** (Google)

2. ***How*** Do Companies Talk About AI Safety?

Metaphors

"Suppose a cleaning robot maker set the objective to remove visible dirt as fast as possible. If the optimal approach turned out to be hiding dirt under the carpet, or throwing away all visible dirty objects, this would be **a failure in spirit even though it might satisfy the objective**" (Google)

2. ***How*** Do Companies Talk About AI Safety?

Metaphors

"Suppose a cleaning robot maker set the objective to remove visible dirt as fast as possible. If the optimal approach turned out to be hiding dirt under the carpet, or throwing away all visible dirty objects, this would be **a failure in spirit even though it might satisfy the objective**" (Google)

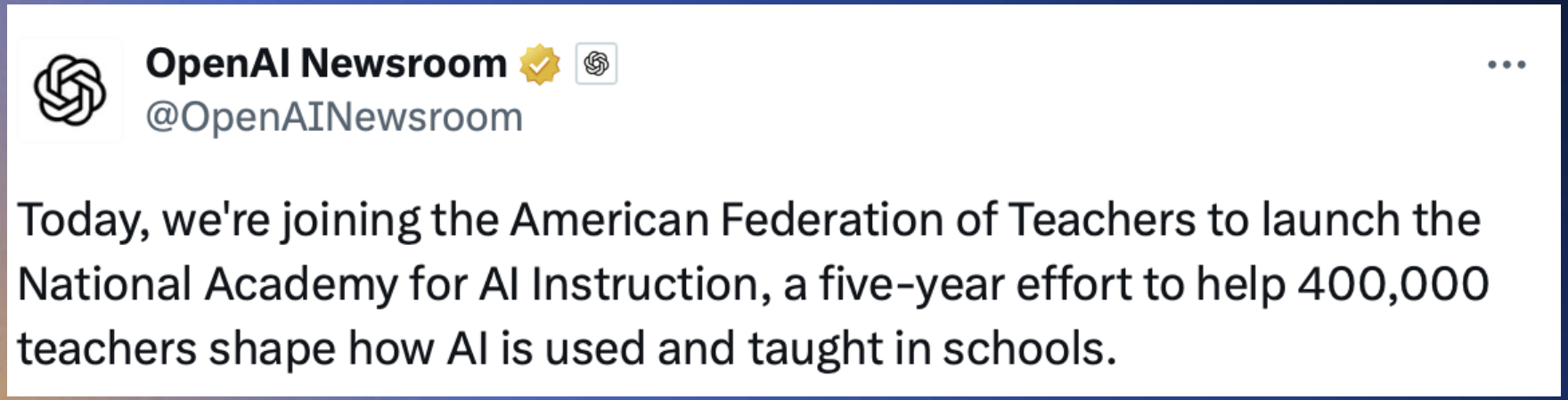
"**Nuclear power stations** have continuous radiation monitoring and regular site inspections; **new aircraft undergo extensive flight tests** to prove their airworthiness. It's **no different for AI systems**..." (**Anthropic**)

Implications for AI Literacy (Emergent)

Dimension	Examples	AI Literacy Insight
Responsibility & Accountability	Forward-looking commitments, distributed across companies, users, and governments; accountability often vague	Responsibility ≠ enforceable accountability; signals diligence without clear consequences
Governance & Oversight	Internal safety teams, policies, multi-stakeholder collaboration; cautious calls for regulation	Recognize authority construction and regulatory shaping through organizational structures
Risk & Mitigation	Iterative evaluation, testing, red-teaming, monitoring, user feedback	Safety as an ongoing process, not a static property
Dynamic & Global Framing	Safety portrayed as emergent, evolving, and globally consequential	Safety is contingent, context-dependent, and socially negotiated
Metaphors & Analogies	Comparisons to CBRN, nuclear power, aviation; operationalized through everyday scenarios	Connects abstract AI risks to familiar real-world scenarios, highlighting both constraints and potential misunderstandings of AI's impact relative to existing technologies

Discursive Toolkits: Operationalize interpretive and critical competencies

Implications for AI Literacy (Emergent)



Discursive Toolkits: Operationalize interpretive and critical competencies

Implications for AI Governance (Emergent)

NEWS AI POLICY

Europe is scaling back its landmark privacy and AI laws

Image: The Verge

/ The EU folds under Big Tech's pressure.

by Robert Hart and Dominic Preston
Nov 19, 2025 at 1:47 PM GMT+1

61 Comments (All New)

Implications for AI Governance (Emergent)

- Participatory language is a powerful rhetoric
  - But it also distributes responsibility
  - Creates leeway for “mistakes”
  - Brings in an innovation-first language (“surgical?”)

Governance at a distance: “**to navigate the complexities of the present, it is deemed better to aim for a horizon that is as remote as possible and hope for the best**” (Roberge et al., 2020)

“AI is one of the most important things humanity is working on..”

Sundar Pichai, CEO of Google, 2018

then how we talk about it, define it, and govern it, matters !!!

Literacy: Discursive competencies for AI literacy

Governance: Separation between stakeholders, impact of discourse on regulation, should discourse on such an important technology be regulated?

Thank you

Do you have any questions?

Ankolika De, PhD Candidate, Penn State;
apd5873@psu.edu

Also, on behalf of **Gabriel Lima and Yixin Zou**