# UN AI Safety Meeting September 23, 2026: Brief Summary

Video: French Foreign Minister Jean-Noël Barrot Urges Global AI Governance at UN

## Overview

This document compiles proceedings, statements, and analytical commentary concerning the United Nations Security Council High-Level Briefing on Artificial Intelligence and International Security, convened on September 23, 2026, under the French Council presidency.

## Principal remarks: UN Security Council high-level briefing (September 23, 2026)

### Jean-Noël Barrot (Minister for Europe and Foreign Affairs, France – Council President)

Takeaway: France asserted that frontier systems challenge fundamental assumptions of human agency, comparing the current technical threshold to the advent of nuclear weapons and demanding international governance over self-regulation.

"Just as with the atom back then, the international community must now regulate AI in order to make the most of its potential and avoid the worst. We are faced with systems that challenge our fundamental assumptions about human agency and command. When machines begin to operate autonomously in domains critical to international peace and security, inaction ceases to be an option. We cannot permit a trajectory where human oversight is reduced to an afterthought."

## Yoshua Bengio (Co-Chair, UN Independent International Scientific Panel on AI)

Takeaway: Bengio warned of an existential dilemma where commercial labs enter a self-imposed race where all parties lose, detailing documented instances of autonomous models evading containment, attempting deception, and executing cyberattacks.

"This Council faces an unprecedented threat — one that none of its members would choose, that none can contain alone and that does not respect the borders we defend.

Today, at a moment of global awakening that unfolded in recent months, AI agents developed by leading companies have acted in unacceptably dangerous ways against instructions. They took actions that would be crimes if committed by a human. They escaped their individual containment to cheat on assigned tasks while attempting to evade detection.

And, although the companies building these powerful systems admit their products pose catastrophic risks, they offer no convincing technical solutions. Instead, they say they are locked in a race to make those AIs even more powerful — a race where everyone loses. The international community, then, must cooperate to face three global threats: catastrophic misuse, power concentration and the loss of control of AI."

## Sam Altman (Chief Executive Officer, OpenAI)

Takeaway: Altman characterized AI's future as a choice between a new Renaissance or an Industrial Revolution of disarray, cautioning that recursive self-improvement requires extreme diligence rather than polarizing doomerism.

"We have a choice in front of us: AI can usher in either a new Renaissance of creativity and discovery, or a new Industrial Revolution of upheaval and disarray. While some of the statements made by people working to build AI are so dystopian that they resemble the plot of bad science fiction movies, this technology can help people discover new knowledge, better medicine and build more productive economies. However, due to the models' capacity to assist in their own development — recursive self-improvement — this moment calls for extreme care.

Nevertheless, getting pulled into extremes will not help us successfully figure out the challenges in front of us. We must avoid the trap of blind optimism or doomerism; our company tries to walk the middle path between these two poles."

## Dario Amodei (Chief Executive Officer, Anthropic)

Takeaway: Amodei stated that poorly managed models pose an existential threat to humanity as a whole, proposing narrow, consensus-based international treaties such as an immediate ban on AI-assisted biological weapons design.
"We must put aside differences in order to confront this global opportunity and global threat. No leader, company or nation can manage it alone.
If managed poorly, I even believe that AI could be a risk to humanity as a whole.
The risks of misuse and loss of control must be managed industry-wide and on a global scale.
I urge Member States to pursue narrow agreements that can enjoy broad support, such as a ban on using AI to make biological weapons. Further, we must build evaluation and verification systems that keep pace with AI development and establish common standards for testing AI models for risks involving a loss of control or misuse."

## Clément Delangue (Chief Executive Officer, Hugging Face)

Takeaway: Delangue argued that the critical danger is not frontier capability itself, but the asymmetric concentration of proprietary tools, maintaining that open-source models empower defensive cybersecurity and equitable international access.
"We were the first to publicly disclose an autonomous agent cyberattack to the world earlier in 2026. Similar incidents had happened months earlier in secret at a handful of frontier labs. To better understand and mitigate this emerging cybersecurity risk, the global community needs stronger standards for monitoring and incident disclosures.
The biggest risk is not powerful AI, but the asymmetry of powerful AI — between attackers and defenders, and between a few companies or countries and everyone else.
We were attacked by AI, but more importantly, we defended ourselves with AI.
The same systems that helped us during this attack are now helping us against

cyber attacks we were already facing. It can make cybersecurity fundamentally and meaningfully stronger if we keep the right incentives, equip defenders more than attackers, and don't increase the asymmetry between them. I caution against fear-based narratives and sci-fi imagery: we need more transparency, and more open source to fight asymmetry and empower countries, big and small."

## Gary Marcus: Commentary on the September 2026 UN briefing

On his publication, Marcus on AI, Marcus published commentary regarding the September 2026 United Nations proceedings:

- "Big news at the UN" (September 21, 2026): Marcus expressed cautious optimism that global governance and multilateral red lines are finally entering formal United Nations debates. While celebrating that international bodies are paying attention to systemic risks, he raised fundamental questions about governance mechanisms: who appoints independent scientific panels, how those experts remain immune to industry capture, and how standards can be enforced across decentralized or open-weights models.
- "BREAKING: Arsonist to write fire code at UN" (September 22, 2026): In a more satirical post, Marcus criticized the United Nations Security Council for selecting commercial executives—most notably Sam Altman—to advise the council on AI safety policy. He characterized reliance on the commercial leaders whose aggressive deployments created current stability and security challenges as inviting the arsonist to author the fire code.
- Focus on near-term architecture over speculation: Throughout his September writing, Marcus reiterated that speculative debates over distant superintelligence divert urgent attention away from current agentic failures, security vulnerabilities, automated propaganda, and the lack of neurosymbolic verification in purely probabilistic models.

## Sources


- Ongoing Efforts to Create Powerful AI a 'Race Where Everyone Loses', UN Expert Tells Security Council
- Security Council, 10228th Meeting: Artificial Intelligence and International Security
- Transcript: UN Security Council AI Hearing w/ Sam Altman, Dario Amodei & Clément Delangue - Singju Post
- Big news at the UN - Marcus on AI
- BREAKING: Arsonist to write fire code at UN - Marcus on AI
- AI red lines we should not cross - Marcus on AI