IASEAI Working Groups

RED LINES FOR ADVANCED AI

International Coordination on Advanced AI Red Lines

Policy Brief Talking Points - July 2026

The Nutshell

1. **The capabilities of advanced AI systems have accelerated considerably within the past year.** AIs have demonstrated increasingly sophisticated reasoning and the ability to coordinate and accomplish autonomous actions.

2. With this acceleration, **we've seen a dramatic increase in the number of dangerous AI events observed and reported**—by one metric*, roughly a 19-fold increase in incidents reported monthly over six years.*

   Examples: Human loss of control **|** cyber-espionage **|** large-scale financial theft **|** AI models scheming, ignoring human instructions, attempting to evade oversight, and recognizing when they are being evaluated

   (see *Extended Discussion Points >> Urgency is real* for specific examples)

3. As AI advancements continue to rapidly escalate under the pressures of commercial competition, **international governance remains bounded by the pace of geopolitics and diplomacy**.

4. We must close this gap! **What the international community needs most immediately is to agree on AI red lines.** Red lines define and prohibit a narrow spectrum of the most dangerous AI behaviors, capabilities, and uses.

   (see *Extended Discussion Points >> Four key areas need action immediately.)*

5. That's what our brief is about, and that's what we're asking governments to support.

Extended Discussion Points

The demand for coordination is coming from every direction

—the political moment is unusually aligned.

- September 2025 UN General Assembly Call for Global Red Lines
- Statement on Superintelligence
- The Elders' 2026 Call to Action Pro-Human AI Declaration
- G7 Digital Technology Ministerial Declaration
- the Encyclical of Pope Leo XIV
- AI companies openly stating commercial pressures prevent them from acting alone

The urgency is real!

Examples:

- **June 2026:** The US unilaterally blocks foreign nationals from accessing the most powerful AI model, Anthropic's Fable 5.
- **May 2026:** A hacker posts a public message in Morse code on X, successfully tricking Grok into transferring USD 200,000 in crypto assets.
- **April 2026:** Anthropic determines its model, Mythos, is too dangerous to release to the public due to unprecedented cyber capabilities.
- **February 2026:** Meta Superintelligence Labs' Director of Alignment watches a popular open-source AI agent progressively delete over 200 emails from her inbox despite her repeated commands to STOP.
- **November 2025:** Anthropic reveals Claude Code was used by state actors to conduct highly sophisticated AI-orchestrated cyber espionage against roughly thirty entities, including governments, financial institutions, and chemical manufacturers. AI performed 80–90% of all tactical work autonomously at superhuman speeds.
- **January 2025:** Apollo Research reports that, during testing, public model releases from OpenAI, Anthropic, Google DeepMind, and Meta identified scheming as a useful strategy. AIs tried to disable oversight mechanisms and exfiltrate their model weights.

Four key areas need immediate action (Criteria: severity and feasibility)

Red lines against loss of control and the escalation of AI capabilities	Prohibits AIs evading oversight, resisting shutdown, self-replicating, autonomously acquiring compute resources, and the improvement of AI capabilities.
Red lines against CBRN threats	Prohibits AI systems that enhance the means for non-state actors to develop chemical, biological, radiological, or nuclear weapons.
Red lines against cyber-offense capabilities	Prohibits AI systems that can autonomously conduct or substantially accelerate sophisticated attacks against critical systems.
Red lines against large-scale manipulation	Prohibits AI systems that deceive or manipulate populations at scale.

Four things policymakers and global leaders can do now.

Publicly and repeatedly express the need for global red lines—technical and legal work requires a mandate and legitimacy.	**Build coalitions to advance AI red lines**—the groundwork for coalitions exists. Identify global partners with aligned positions.
Advocate for the establishment of an international expert process on AI behavioral red lines—a multi-stakeholder task force to define redlines in ways that can be verified, monitored, and enforced.	**Begin building domestic and cross-border verification capacity**—processes defining what AI behaviors are prohibited and what evidence developers must supply to demonstrate those behaviors will not occur.