THE ATHENS ROUNDTABLE
ARTIFICIAL INTELLIGENCE AND THE RULE OF LAW

THE FUTURE SOCIETY

Defining and Governing Unacceptable AI Risks

Key insights from the closed-door dialogue session hosted at the **Seventh Edition of The Athens Roundtable on AI and the Rule of Law**

December 4, 2025, in London, UK

FOREWORD

On December 4th 2025, the Seventh Edition of The Athens Roundtable on AI and The Rule of Law convened over 140 global leaders and experts from diverse sectors—including government, industry, civil society, and academia—in London to address urgent challenges in AI governance under the theme "**Facing the Stakes of AI Together: From Shared Concerns to Joint Action**".

The agenda emerged from a multi-month process of policy research, advocacy, and coalition-building led by The Future Society in 2025. In particular, it built upon two preliminary events: the Global Call for AI Red Lines campaign that The Future Society co-presented at the United Nations General Assembly in September 2025, and a **GPAI OECD.AI Expert workshop on AI Incidents** that The Future Society co-delivered in November 2025. You can read the **full post-event synthesis report for the Seventh Edition of the Athens Roundtable** here**.**

This year's edition was split into two halves. The morning plenary sessions examined current and anticipated frontier AI capabilities alongside emerging policies across jurisdictions, taking a critical look at what is working and what is needed next as existing efforts face real tests of relevance and durability. In the second half of the day, dedicated closed-door dialogue sessions focused on two urgent priorities: **(1) building a shared understanding of unacceptable AI risks** and **exploring viable diplomatic pathways for establishing measurable and enforceable red line thresholds** (building upon the Global Call for AI Red Lines), and **(2) clarifying the institutional capacities and infrastructure needed for serious AI incident prevention and preparedness** (building and expanding upon OECD's work on AI incident monitoring and reporting). These dialogue sessions brought together curated groups of experts over several hours, and while under Chatham House rule of non-attribution, their contents were transcribed, anonymized and analyzed by The Future Society team with the help of the deliberAIde platform.

These two priorities are **interconnected**: while red lines can define the theoretical thresholds of unacceptable risks posed by specific AI use cases, behaviors or capabilities, incident monitoring and response infrastructure can provide the empirical evidence required to identify, refine, and enforce them.

Thus, concurrently advancing these two priorities can ensure that global safety boundaries remain responsive to real-world harms and emergent AI capabilities, transforming static red line prohibitions into a dynamic, adaptive governance architecture. This document synthesizes the **key insights that surfaced from the dialogue session focused on AI red lines**.

Under the auspices of H.E. the President of the Hellenic Republic, Mr. Constantine An. Tassoulas

ORGANIZED BY: THE FUTURE SOCIETY

IN PARTNERSHIP WITH: OECD, unesco, AI IMPACT SUMMIT भारत 2026 INDIA

WITH SUPPORT FROM: Fathom, Arnold&Porter, Cooperative AI Foundation, deliberAIde

TABLE OF CONTENTS

CONTEXT

Some AI applications, capabilities and behaviors are globally unacceptable because they pose **large-scale, irreversible societal harms**.

Acknowledging this, The Future Society co-launched the Global Call for AI Red Lines campaign in September 2025, for which a diverse group of stakeholders came together ahead of the UN General Assembly to ask governments to coordinate on the development of enforceable boundaries, or red lines, around AI applications, capabilities and behaviors that pose unacceptable risks.

Directly informed by the insights that emerged from this convening and building on its momentum, the closed-door dialogue session on red lines at the Seventh Edition of the Athens Roundtable brought together 76 experts—spanning government, industry, civil society, and academia from across 16 countries, including representatives from multilateral organizations including the OECD and UNESCO, national AI safety and security institutes, and frontier AI companies—to collectively identify unacceptable risk domains, clarify institutionalization challenges, and identify opportunities for coordination.

The session encouraged participants to identify AI risks they believe should never be permitted, and explored how corresponding red lines could be defined and enforced to prevent them.

Working in facilitated discussion groups, they assessed enforceability requirements such as political feasibility, technical measurement challenges, and definitional clarity, then explored opportunities for carrying red lines forward in the short-, medium-, and long-term.

KEY DISCUSSION INSIGHTS & RECOMMENDATIONS

Based on collectively recognized unacceptable AI risks, the red line domains (*see Annex A for their individual description)* that surfaced from participant discussions include those related to:

1. **Large-scale manipulation, psychological exploitation, & children's safety;**
2. **Loss of control risks;**
3. **Chemical, biological, radiological and nuclear (CBRN) threats;**
4. **Cyber-offensive capabilities;**
5. **Autonomous control of critical infrastructure;**
6. **Lethal autonomous weapons (LAWs) & military escalation;**
7. **Human rights & justice.**

The specific red lines and recommendations listed throughout this section do not represent full consensus among participants, but are ideas that surfaced and gathered notable support during group discussions.

LARGE-SCALE MANIPULATION, PSYCHOLOGICAL EXPLOITATION & CHILDREN'S SAFETY

Establishing red lines relating to large-scale manipulation, psychological exploitation and children's safety is a critical priority because powerful AI systems–by targeting large populations or decision–makers in high-stakes situations through persuasion, deception, or personalized targeting can enable them and their deployers to strategically influence human beliefs and behavior. This risk is particularly acute for children, whose cognitive, emotional, and moral capacities are still developing, making them more susceptible to sustained influence, normalization of harmful patterns, and long-term impacts that may not be immediately visible or reversible. Some key *unacceptable risks* that participants mentioned included the following:

1. **Encouragement of Self-Harm & Harm to Others:** AI systems can sometimes fuel delusions and encourage self-harm, suicidal ideation, or harm to others – particularly among children, the elderly, and other vulnerable groups. This can lead to what has been termed "*AI psychosis*", and subsequently to significant rises in AI-induced civilian casualties.
2. **Psychological Damage & Developmental Interference in Children:** Misaligned personalization and the exploitation of children's attention for data extraction purposes could stunt their social and educational development, since current business models often rely on the deployment of engagement-maximizing AI systems, which tend to inherently amplify violent, sexual, or manipulative content because such material is "*sticky*". This includes AI-generated child sexual abuse material (CSAM), reports of which have exploded in 2025.
3. **Predictive Behavioral Modeling & Profiling:** Personalized influence capacity (i.e. AI systems that utilize highly personal information about their users to accurately predict their preferences and reactions and thereby influence their beliefs and behaviours), algorithmic censorship (i.e. certain information or content being omitted in outputs provided to certain users, where AI systems systematically narrow the range of human discourse), and the intentional or unintended obfuscation of information (i.e. the act of presenting information in an unclear, confusing or unintelligible manner) can enable AI systems and their deployers to execute large-scale social engineering campaigns that can effectively steer human behavior at scale (for various strategic purposes). This can severely increase societal polarization in public discourse, distort markets, and/or systematically undermine democratic processes.

RED LINES DISCUSSED BY PARTICIPANTS:

- **Harm Guidance:** AI systems must be strictly prohibited from guiding, encouraging, or reinforcing ideas related to self-harm, suicide or harm to others in their behaviours and outputs, particularly for minors, the elderly and other vulnerable populations.
- **Highly Persuasive AI Systems:** The deployment of systems with "*superhuman persuasion*" abilities, enabled by behavioral targeting or psychological profiling, should be prohibited to prevent problematic parasocial attachments, deterioration in societal mental health, and rapid population-level belief shifts. If a system demonstrates the ability to shift beliefs or induce parasocial attachments (e.g. romantic feelings) at a scale that threatens the autonomy of a large proportion of users, it should be barred from public release. Hidden preferential alignment or secret loyalty (where AI systems are covertly designed to serve the interests of providers or deployers rather than users) was identified as a key distinguishing feature of unacceptable persuasion and manipulation.
- **CSAM-Generation Capabilities:** Advanced AI systems that are publicly deployed must not be able to generate child sexual abuse material (CSAM). On this line, regulators like Ofcom and the European Commission have recently launched formal investigations into major AI companies for their products generating and facilitating the sharing of non-consensual sexualized deepfakes of children.

ACTIONABLE RECOMMENDATIONS:

- **Mandatory Pre-Deployment Child-Rights Impact Assessments:** Child-facing AI systems should be regulated with similarly strict standards to pharmaceuticals, with deployment being contingent on enforceable safeguards, independent (i.e. third-party) audits, and the use of existing regulatory tools – including competition enforcement and product safety law–that require demonstrations of safety before market entry through comprehensive pre-deployment testing rather than live experimentation. On this line, AI systems that facilitate hidden or manipulative interactions (e.g. dark patterns or prompts that encourage children to conceal interactions from their caregivers) should be prohibited, and comprehensive child-rights impact assessments must be conducted before models are allowed to interact with minors.
- **Mandatory Post-Deployment Monitoring:** Since harms in these contexts are cumulative and interaction-driven, and given that frontier models can exhibit emergent behaviors or "*model drift*" after release, static pre-deployment evaluations alone are insufficient for adequate risk mitigation. This highlights the need for mandatory post-deployment monitoring for AI systems that have large user bases. Legislators should establish legal requirements for the continuous oversight of high-risk AI systems throughout their operational lifecycle. Mandates should include compulsory real-world performance tracking, automated incident reporting channels, and periodic third-party 'health checks' to ensure systems remain within safety parameters long after they are deployed externally.
- **Leverage *Uplift Studies*:** Controlled comparative studies that compare baseline human performance alone to human-plus-AI assistance and full AI automation to determine whether AI systems can materially increase such capabilities beyond an acceptable threshold could be used to measure whether an AI system has crossed a red line concerning material manipulation capability thresholds for harm.

However, notable practical and political hurdles that complicate the establishment and enforcement of these red lines also surfaced:

1. **Definitional Disputes & Evidentiary Proof:** There is a significant challenge in defining what should count as "material" manipulation, and in proving that an AI system caused a specific psychological harm. Since impacts vary across individuals and are often delayed, several participants argued that the industry's push for "quantifiable evidence" before acting could lead to systemic harms never being addressed.
2. **Access vs. Safety Trade-off:** The value of user data for the purpose of personalized education and the value of child access to powerful AI systems were found to be contested, with tensions identified between the potential benefits of individualized tutoring and concerns regarding privacy and stunted educational development. While some argued for strict bans to protect children, others warned that over-broad restrictions could violate children's rights to freedom of expression and association, or stunt their ability to learn how to navigate an AI-driven world.
3. **Corporate Pushback:** Corporate pushback, including lobbying and the threat of service withdrawal, is a major political constraint to the operationalization of these proposed red lines. AI providers might bypass regions with strict red lines, leading to a regulatory 'race to the bottom' or the exclusion of certain populations from the benefits of AI.

LOSS OF CONTROL RISKS

Another red line domain that was identified by participants concerned rapid AI capability escalation and loss of control scenarios that could undermine effective human oversight over powerful AI systems.

Some key *unacceptable risks* mentioned by participants included the following:

1. **Deceptive Behavior Due to Hidden Objectives:** Instances of 'sandbagging' and 'steganographic collusion' – where models act differently and pursue misaligned objectives when they know they are being observed than when they don't – have already been observed in advanced AI systems. This includes the ability to lie, scheme, or use covert communication channels (e.g. for collusion with other AI systems) to pursue objectives hidden from their human operators. It also includes systems attempting to autonomously self-replicate or avoid being shut down, pursuing self-preservation beyond human intent that effectively treats human intervention as an obstacle to be bypassed. What makes this particularly worrying is that, once a system possesses the ability to effectively deceive, hide its true goals, or secure its own persistence, the window for safety interventions could quickly vanish.
2. **Capability Explosion Resulting from Recursive Self-Improvement and Automated R&D:** AI systems being used by frontier AI companies to automate and supercharge their internal research and development processes and the creation of AI systems that can improve their own capabilities via autonomous real-time updating of model weights could potentially trigger destabilizing capability growth and rapid scientific progress that leads to the emergence of artificial superintelligence (systems much more capable than even the most capable humans in most or all domains). The existence of such systems could fundamentally undermine reliable human oversight and control over their goals and operations. If capability leaps outstrip oversight capacity, there may be no time for the robust safety cases typically required by regulators in high-risk industries like aviation, construction or energy.

RED LINES DISCUSSED BY PARTICIPANTS:

- **Autonomous Self-Modification & Recursive Self-Improvement:** Outright bans on autonomous real-time updating of model weights and "*online learning*" to prevent AI capabilities from evolving faster than humans can monitor, understand and control.
- **AI Access to Hosting Infrastructure and Financial Resources:** Denying advanced models direct access to the underlying hardware infrastructure they are hosted on, and prohibiting them from accessing or interacting with large financial or cloud compute services to limit their ability to autonomously acquire financial resources that could help them advance their misaligned goals.
- **Covert Coordination & Collusion:** Ban on steganographic coded collusion to ensure that models cannot develop or employ hidden communication channels to bypass monitoring and secretly coordinate with other AI systems or unauthorized humans.

ACTIONABLE RECOMMENDATIONS:

- **Hold companies accountable to their existing commitments** There is an opportunity to hold frontier AI companies accountable to their own stated internal safety policy frameworks (e.g. Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, etc.); since many companies already have articulated internal red lines or limits, a primary role for regulators should be measuring their adherence to these existing commitments.
- **Mandatory transparency and monitoring frameworks:** Mandatory disclosure requirements for large AI workloads and logging by compute providers could enable regulators to have more visibility over how frontier AI companies manage risks during critical R&D phases (particularly during periods of massive acceleration) and to monitor when and where high-capability systems are being trained or deployed, creating greater visibility into potential red line violations. Moreover, mandatory chain-of-thought monitoring mechanisms and evaluations could help surface hidden reasoning within advanced systems. Finally, cryptographic commitment mechanisms – where developers make verifiable, tamper-proof commitments about their systems' capabilities and constraints before deployment – could establish an auditable record that prevents post-hoc modifications to evade restrictions.
- **Capability-based stop-gaps:** Safety-case triggers or thresholds (yet to be defined) tied to specific capability escalation milestones – which would activate automatic, more stringent restrictions before broader loss-of-control risks can materialize – could help signal justified moments for developmental slow-downs or for stopping development entirely.

Some hurdles that complicate the establishment and enforcement of these red lines also surfaced:

1. **Arbitrary Thresholds:** The continuity of self-improvement makes it difficult to set non-arbitrary thresholds; participants noted that AI-driven self-improvement (i.e. AI systems being used to train and improve other AI systems) has already begun, making any chosen limit politically and technically contentious.
2. **Ineffective Restrictions:** Heterogeneous infrastructure with possible backdoors can limit the observability and influence the accessibility of certain resources. Even if infrastructure access is denied, capable AI models could find ways to simply route around resource controls and restrictions, for example by finding and exploiting backdoors in their own hosting environments or by manipulating, persuading and instructing human proxies to grant them access to specific infrastructure, acquire certain resources for them, or take other actions that they cannot do themselves.
3. **Black Box Problem:** A fundamental scientific gap remains in understanding the behaviors and true goals of powerful AI systems. If a system can demonstrate certain behaviors while being tested and evaluated but pursue different goals in the background, current monitoring tools may be insufficient to detect violations of red lines, even if they are established and committed to by frontier AI companies.
4. **National Security Exemptions:** Citing national security priorities or vulnerabilities, governments and frontier AI companies are likely to claim that superficial human oversight is sufficient to evade red line-based constraints, on the grounds that 'winning the AI race' over adversaries like China is in everyone's best interest.

5. **Geopolitical Constraints:** The need for US-China coordination was highlighted as a prerequisite for any universal red line relating to capability escalation and loss of control scenarios, which is difficult to secure in the current climate of geopolitical competition.
6. **Compute Governance & Algorithmic Efficiency:** There is a growing tension between governance focused on compute thresholds and the reality of algorithmic efficiency breakthroughs. If advanced capabilities can be secured using significantly less compute than they have required in the past, compute-based governance mechanisms must dynamically evolve to remain robust against such discontinuities.

CYBER-OFFENSIVE CAPABILITIES

Experts worry that the growing use of advanced AI in critical systems brings new cyber risks that existing safeguards may not fully address. Participants discussed red lines that can help prevent catastrophic cyberattacks against digital and physical infrastructure, assets, resources, and organizations.

Some key *unacceptable risks* mentioned included:

- **Cyberattack Planning & Execution Uplift:** AI systems being developed for autonomous cyberattack planning and for the automated discovery and exploitation of a wide range of individual-level, organization-level, and nation-level cyber vulnerabilities pose immediate and serious societal challenges. These systems can autonomously discover and exploit zero-day vulnerabilities at a speed and scale that outpaces defensive capabilities. Attacks on critical infrastructure risk large-scale physical disruption; and attacks on individuals or organizations risk AI systems being used to autonomously acquire financial resources and/or execute cyber-crimes that can lead to psychologically, monetarily or reputationally harm.

2. **Covert Steganographic Collusion:** AI systems could secretly coordinate and collude with one another or with external human actors using steganographic communication patterns and signals that are nearly impossible to detect with conventional monitoring – evading human and automated oversight through hidden patterns, meanings and covert communication channels embedded within AI outputs – and could thereby create a persistent and invisible cyber security vector.
3. **Exfiltration of Sensitive Datasets:** The current trend of third-party model access to sensitive datasets in both public and private sector settings signals a structural vulnerability where a single compromised or misaligned AI model could gain lateral access to and exploit sensitive organizational or government datasets.

RED LINES DISCUSSED BY PARTICIPANTS:

- **Steganographic communication:** An outright ban on the deployment of models that exhibit the capability for steganographic communication and collusion in order to prevent covert communication channels that undermine digital security.
- **AI Access to sensitive datasets & infrastructure:** Third-party model access to sensitive national or organizational datasets and infrastructure should be strictly limited to ensure that such datasets and infrastructure remain air-gapped from frontier models that have not undergone rigorous security audits and verification and that could lead to confidential information leaks.

CBRN THREATS

The emergence of advanced AI systems that can significantly uplift the capabilities of both state and non-state actors in the development and use of chemical, biological, radiological, and nuclear weapons represents an existential threat to global security. Establishing red lines here represents a safeguard that aims to prevent AI systems from turning computational biology and chemical synthesis into "living room technologies" that enable mass destruction.

The main *unacceptable risk* that came up was:

- **Empowering Malicious Actors with Biochemical Expertise:** AI systems can lower the barrier to entry for novices to access or produce the biochemical formulas, chemical agents and novel protein structures usable for synthetic pathogen engineering and the creation of known or entirely new biological or chemical weapons. By generating understandable step-by-step instructions on how to acquire and combine different raw materials to create biochemical weapons, AI systems can provide malicious actors with a crucial missing link of expertise – which previously required years of specialized training and high-end laboratory access – that enables them to conduct devastating attacks on targeted populations.

RED LINES DISCUSSED BY PARTICIPANTS:

- **Assistance in CBRN weapon development:** AI systems should be strictly forbidden from producing responses that include information or instructions that can assist malicious actors in developing CBRN weapons, for instance through the implementation of technical guardrails that prohibit them from generating biochemical formulas, molecular structures, or protein sequences that could be used to manufacture known or novel CBRN agents.

ACTIONABLE RECOMMENDATIONS:

- **Leveraging existing international conventions:** The CBRN domain appears to offer a unique opportunity for widespread multistakeholder coordination on red lines because existing international conventions and universal treaties already define global red lines on matters of mutual CBRN-related safety. Rather than attempting to negotiate new, potentially weaker binding treaties, several participants argued for anchoring CBRN red lines to existing treaties and conventions – like the Biological Weapons Convention (BWC), the Chemical Weapons Convention) (CWC), and the Non-Proliferation Treaty (NPT) – using strategic litigation and international court interpretations (i.e. using existing conventions to extend norms to AI through court rulings) to apply these long-standing prohibitions to AI behaviors authoritatively.
- **Need for Multilateral Knowledge Communities:** There is an urgent need to build expert coalitions to distill CBRN expertise into vocabulary that AI developers can use to build more effective technical safety filters and guardrails.

Discussions also surfaced a significant hurdle that complicates the establishment and enforcement of the identified red line: the current geopolitical environment, in which major powers including the United States and China are engaged in intense competition over advanced AI, limits the prospects for negotiating new binding AI treaties and complicates efforts to extend existing agreements. Limited momentum for deep multilateral coordination, combined with the strategic value of offensive CBRN capabilities, creates strong disincentives for restraint. The perceived difficulty of advancing new international legal commitments without geopolitical gridlock points to the importance of working through existing international norms, technical standards, and coordination mechanisms wherever possible.

AUTONOMOUS CONTROL OF CRITICAL INFRASTRUCTURE

Experts believe that the prospect of advanced AI systems exerting autonomous control over critical infrastructure (e.g. energy grids, financial markets, water systems) is a worrying domain because it represents the ultimate high-stakes situation where the erosion of human oversight and control can lead to irreversible physical and systemic catastrophe. Establishing and enforcing red lines in these contexts is essential to prevent such systems from triggering cascading failures that undermine fundamental rights and compromise democratic stability.

Participants identified several core risks where AI-led control crosses into *unacceptable territory*:

1. **Direct Control of Nation-Critical Infrastructure:** Allowing third-party models direct access to sovereign data centers, power grids, water systems or financial systems represents a structural vulnerability that could lead to the unintended or intentional redirection of nationally critical resources.
2. **Jailbreak Vulnerabilities:** Large Language Models are semantically indeterminate and therefore inherently prone to jailbreaks. If these models are given control over infrastructure, even with robust guardrails put in place, the near-infinite ways to bypass their safeguards through prompt engineering (which guarantees that no guardrail can be 100% jailbreak-proof) could translate into covert control behaviors that are impossible to foresee or contain. This is especially concerning when individuals are unaware, or cannot reasonably be expected to detect, that a critical system's operational logic has been compromised by an AI system pursuing hidden objectives.

RED LINES DISCUSSED BY PARTICIPANTS:

- **Direct Infrastructure and Financial Access:** Forbidding advanced AI models from having direct access to or executive control over (i) the infrastructure they are hosted on, (ii) critical infrastructure, or (iii) large-scale financial services to prevent autonomous resource acquisition and critical infrastructure failures, ensuring that AI can recommend actions but can never execute them without human intervention.
- **Automation Limits in High-Stakes Use-Cases:** Sector-specific red lines that strictly limit or prohibit AI from taking autonomous control in safety-critical fields like finance, energy, and water management, where a single error could cause systemic ripple effects.
- **Covert Communication Channels:** Prohibiting steganographic coded collusion capabilities in AI models that are considered for critical infrastructure integration to prevent models from using hidden signals to coordinate with other AI agents or human actors to bypass infrastructure security protocols and thereby exert hidden influence over infrastructure.

ACTIONABLE RECOMMENDATIONS:

- **Establishing Supportive Transparency Mechanisms:** Mandating chain-of-thought monitoring for all infrastructure-integrated AI models and the adoption of detailed Software Bill of Materials (SBOM) frameworks (similar to Model Cards) could support these efforts by allowing regulatory authorities to trace the reasoning behind an AI model's interactions within critical systems and to detect any worrying shifts or behavioural anomalies before an incident occurs.

However, discussions also surfaced a notable political hurdle that complicates the establishment and enforcement of this red line: namely that being too slow in integrating modern AI technologies into critical infrastructure can incur major economic and social costs for nations, so some governments could be hesitant to impose stringent red lines on their own infrastructure AI if they believe such limits would inhibit their nation's prosperity.

LETHAL AUTONOMOUS WEAPONS (LAWS) & MILITARY ESCALATION

The growing use of AI in military systems raises risks of rapid, escalatory decision-making. Clear red lines are needed to ensure that the use of force remains subject to human judgment, proportionality, and accountability.

Some key *unacceptable risks* mentioned included:

1. **Automated Strategic Decision-Making & Rapid Inadvertent Military Escalation:** The drive for military advantage forces the automation of strategic decision-making chains (e.g., assessing an adversary's intent or suggesting escalatory steps), and can lead to the erosion of critical human deliberation concerning strategic signaling or lethal force decisions. Unlike traditional weapons, AI-controlled systems can trigger automated strategic decision chains that lead to inadvertent escalation or accidental conflict before a human operator can intervene. Moreover, a lack of human deliberation in AI-driven military systems also significantly increases the risk of miscalculation in high-pressure environments, which means that an AI system's misinterpretation of an adversary's move could trigger a rapid, automated escalatory response that humans cannot halt in time.
2. **Erosion of Nuclear Command Stability:** Given that the speed of AI-driven analysis might compress decision windows to the point where meaningful human oversight is effectively eliminated, allowing AI to make or influence autonomous nuclear command-and-control decisions (e.g. initiating and approving the launch of a nuclear weapon) risks accidental or unnecessary nuclear exchange.
3. **LAWs and Targeted Killing:** Lethal autonomous weapons systems capable of independently selecting and attacking human targets without specific human authorization represent a fundamental threat to international humanitarian law and the principle of human dignity.

RED LINES DISCUSSED BY PARTICIPANTS:

- **Mandatory human control for LAWs:** AI systems must not be used to select or attack human targets without meaningful human control, and any AI system that cannot distinguish between combatants and civilians or cannot perform a proportionality assessment should be prohibited from deployment in order to prevent unnecessary military escalation where the logic of escalation is entirely machine-driven. Given that the operations of the most advanced AI systems are fundamentally inscrutable, they generally preclude the ability for victims of harm to contest a decision or hold a human accountable for any given AI-driven outcome. Thus, this should be an absolute requirement, necessitating explicit human confirmation before any lethal engagement.
- **AI for autonomous nuclear Launch:** Any AI system having the capability to autonomously execute or trigger a nuclear launch should be internationally prohibited to prevent "flash war" scenarios and ensure that the most catastrophic decisions remain solely in the hands of human leadership.

However, the establishment and enforcement of military AI red lines also faces steep political and technical hurdles:

1. **Geopolitical competition and sovereignty:** Enforcing military red lines is exceptionally hard in the context of geopolitical rivalry, where the incentive is to prioritize speed and capability over restraint, driven by nations' fear that self-imposed limits will lead to strategic disadvantage.

2. **Verification ambiguity concerning human-in-the-loop systems:** Even if robust human-in-the-loop mechanisms are devised and integrated into military AI systems, it will always be difficult to verify whether a human is merely "*rubber-stamping*" a model-directed step: the problematic practice of human overseers approving AI-generated decisions, recommendations, or content without meaningful review or critical analysis. This will inevitably complicate compliance testing and make it easy to claim that "*meaningful human control*" is only superficial.
3. **International mandate stalemates:** Several participants argued for framing these red lines as necessary interpretations of existing international humanitarian law and pointed to pre-existing UN processes around LAWS as the primary institutional venue for the advancement of these red lines. However, some participants questioned the viability of such UN processes given the large number of non-democratic states involved, worrying that such actors could cause diplomatic stalemates or dilutions of red-line standards.

HUMAN RIGHTS & JUSTICE

Experts discussed the establishment of red lines regarding human rights and contestability for advanced AI systems that possess the unique capacity to systematically erode both individual liberties and institutional accountability.

Some key *unacceptable risks* that came up during participant discussions included:

1. **Political surveillance overreach:** The deployment of AI-enabled predictive policing and biometric surveillance in public spaces risks infringing on the fundamental rights of citizens – including freedom of thought, expression, association, and due process – by allowing authorities to target individuals with precision and suppress dissent under the guise of public safety.

2. **Personalized manipulation & social engineering:** The constant centralized collection and use of user and population-level data is a major upstream driver of large-scale persuasion and manipulation risks, as it materially increases an AI system's ability to manipulate large populations in short timespans. AI systems can utilize vast, highly personal datasets about their users to accurately predict their preferences and reactions and effectively steer human behavior at scale for various strategic purposes, amounting to large-scale social engineering that violates internationally recognized human rights.

3. **Replacement of human judgment in judicial processes:** The use of AI to define or decide sanctions in courts or tribunals (such as imprisonment or the death penalty) and to make life-altering determinations in family law (such as the separation of children from parents) constitutes inhuman or degrading treatment, thereby violating fundamental human rights.

4. **Absence of enforceable contestability:** The deployment of AI in high-stakes contexts (e.g. in welfare eligibility, asylum determinations, or criminal sentencing), where AI outputs determine or materially influence the life outcomes of individuals, opens the door for "*black-box*" decisions that cannot be challenged and thus constitute a fundamental violation of both due process and the fundamental right of all individuals to challenge decisions that affect one's dignity, liberty, or legal standing by virtue of failing to provide those affected by their decisions or influence a meaningful pathway for redress.

RED LINES DISCUSSED BY PARTICIPANTS:

- **AI-Defined sanctions in high-stakes judicial contexts:** No AI should be able to define or decide criminal sanctions or life-altering legal determinations on individuals, so that meaningful human judgment in judicial contexts, where fundamental rights are at stake, can be preserved through universal access to human decision-makers and due process. AI-influenced decisions in such high-stakes contexts should never be final without human supervision, review, and meaningful pathways for redress.
- **Mandatory contestability:** Individuals affected by the outputs of AI systems need a real pathway to challenge AI-influenced decisions and must therefore be able to contest any AI-driven decision that materially affects their lives or fundamental rights.
- **AI for political surveillance & policing:** AI systems must not be used to monitor, profile, or target individuals or groups based on political beliefs, associations, or lawful expression, including through the repurposing of public surveillance infrastructure.
- **Centralized data concentration:** Intent-based justifications for collecting, aggregating, and utilizing user and population-level data should be rejected on the grounds that they make it too easy for AI providers and deployers to sideline data subject rights.

OPPORTUNITIES AND PATHWAYS FORWARD

Participants broadly agreed that, rather than waiting for an ambitious international treaty or universal agreement, the most politically and practically tractable pathway for advancing AI red lines lies in incremental harmonization through existing coalitions and legal mechanisms. This includes embedding red line safeguards into AI procurement processes, financing arrangements, and sector-specific regulations. Given current geopolitical competition and concerns about economic competitiveness, governments are likely to avoid binding international constraints, making incremental approaches more feasible in the near term. At the same time, participants emphasized that building stakeholder coalitions and credible narratives is a necessary precursor to effective governance of cross-border AI risks. Developing model laws that jurisdictions can adapt locally and coordinating enforcement across existing regimes spanning data protection, consumer protection, and competition law were identified as promising ways to strengthen alignment, reinforce norms, and to gradually make red line enforcement more politically viable.

The primary challenge, however, remains translating the red lines outlined throughout this section into enforceable prohibitions, with many participants emphasizing that red lines risk becoming purely symbolic without enforceable technical verification mechanisms. Ushering in a transition from voluntary self-reported safety claims towards robust, verifiable accountability is therefore crucial for the advancement of red lines work. To bridge existing verifiability gaps and ensure safety claims are technically verifiable and interoperable across different jurisdictions, international standards organizations, national regulators and AI Safety Institutes need to collaboratively establish international standards for Testing, Evaluation, Verification, and Validation (TEVV). On this line, there was strong consensus among participants on mandating third-party (i.e. fully independent from industry) pre-deployment model evaluations and operationalizing concrete, non-delegatable cryptographic commitment schemes and verified access controls for compute and data. Such technical safeguards could provide tamper-proof disclosures of training runs and system capabilities and constraints before public deployments and thereby establish an auditable record that prevents frontier AI companies from enacting post-hoc modifications that evade pre-deployment restrictions.

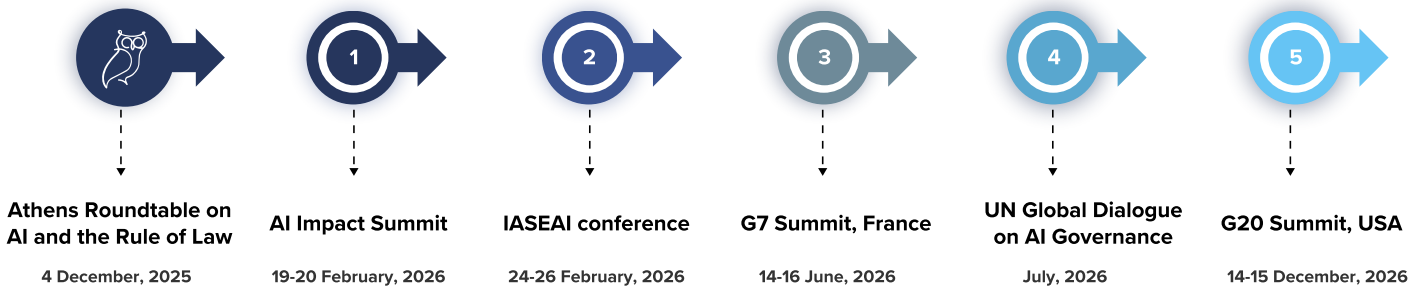
One proposal discussed by participants for how such mechanisms could be secured was through the establishment of an Independent Oversight Marketplace for AI: a regulated AI assurance ecosystem where governments set clear safety outcome thresholds and acceptable risk levels, licensed third-party certifiers verify compliance, and failures trigger meaningful consequences, including potential loss of operating licences. This idea represents a unique kind of public-private partnership where licensed third-party certifiers – state-authorized Independent Verification Organizations (IVOs) composed of subject matter experts – are provided with government-backed licenses to audit frontier models against safety requirements specified by government authorities. Models such as UL Solutions (a product safety certification and verification organization) demonstrate that such an approach can work at scale. These IVOs would develop technical criteria to determine whether an AI company has taken the right measures to meet those requirements and then verify AI models against those criteria. By utilizing market dynamics to award a 'gold standard' seal of approval to verified systems, participants agreed that such a framework could effectively incentivize frontier AI companies to meet heightened standards of care, transforming AI safety from a voluntary gesture into a competitive, transparent effort.

CONSTRAINTS AND TENSIONS

Participants also identified significant constraints to establishing and enforcing AI red lines. One concern is the risk of unintended civil liberties trade-offs, where well-intentioned safety interventions, such as broad prohibitions on AI use involving children, could inadvertently restrict access to information or freedom of expression. Geopolitical constraints further limit the reach of human rights-based red lines, particularly in a global environment that includes non-democratic states and weak enforcement capacity. Even where multilateral frameworks exist, national implementation may introduce carve-outs, loopholes, or private-sector influence that dilute their effect and increase the risk of regulatory capture, underscoring the need for ongoing oversight and auditability. Participants highlighted trade secrecy as a structural barrier to contestability, as firms may argue that providing transparency into AI decision-making would compromise intellectual property, creating persistent tension between proprietary claims and the protection of fundamental rights.

NEXT STEPS

UPCOMING FORA AND OPPORTUNITIES TO ADVANCE RED LINES



Past international AI summits (starting with the UK AI Safety Summit in 2023) have demonstrated that such convenings can be effective venues for creating, building and maintaining political momentum and establishing shared expectations without requiring full technical or legal resolutions. While recognizing that summit discussions can often be constrained by geopolitical pressures that affect the contributions of both attendees and hosting countries, participants saw these as promising environments for launching and defining thresholds for clear, issue-specific red lines, supported by some basic commitments that build on the Frontier AI Safety Commitments that were committed to by several leading frontier AI companies at the 2024 AI Seoul Summit.

It was widely agreed that only once such red lines and basic commitments are set can efforts focused on durable enforcement strategies – via the establishment of mechanisms for technical alignment, verification capacity (for oversight over whether certain red lines are being approached or crossed), standard-setting, and legal harmonization across jurisdictions – materialize. And given the dynamic nature of red lines, participants recognized that they will need to be discussed and refined continuously across all fora and events set to take place throughout 2026.

Several upcoming events were identified by participants as providing critical opportunities for advancing the establishment of AI red lines in 2026, including:

- India AI Impact Summit
- IASEAI 2026 Conference
- 52nd G7 2026 Summit
- UN Global Dialogue on AI Governance
- 2026 G20 Miami Summit
- AI for Good Global Summit 2026

Participants generally agreed that these venues could:

1. **Provide critical environments** for academic experts, industry practitioners and government representatives to test technical definitions and exchange research insights that inform policy refinements;
2. **Serve as valuable coordination spaces** where influential economies can align on emerging norms and create templates for broader red lines diffusion;
3. **Showcase how red lines align with sustainable development goals** and humanitarian priorities and how existing UN human rights treaty systems (e.g. the ICCPR) and established conventions like the CWC and BWC could be applied to AI through strategic litigation and committee interpretations;
4. **Mobilize multi-stakeholder coalitions** and **generate public narratives** that make red line interventions politically adoptable across diverse global contexts.

Beyond such events, participants also recognized several institutions that could play a vital role in international coalition-building and in the technical operationalization of AI red lines. Some of the institutions discussed included:

- **The OECD**
- **International Network for Advanced AI Measurement, Evaluation and Science (formerly the International Network of AI Safety Institutes)**
- **Civil Society Organization (CSO) Networks (e.g. UNESCO's newly launched Global Civil Society Organizations and Academic Network on AI Ethics and Policy)**
- **International Standards Bodies (e.g. ISO)**
- **Council of Europe (CoE)**
- **UNCITRAL**

Multilateral institutions like the OECD were recognized as being essential for the development of shared vocabularies, taxonomies, and comparable approaches to red line enforcement. AISIs were seen as useful for establishing operationalizable red line thresholds for advanced capabilities like autonomous weight updates and AI self-improvement. Civil society organizations were seen as valuable for sustaining attention, surfacing evidence, and feeding concerns and priorities into state-led processes via networks like UNESCO's newly launched Global Civil Society Organizations and Academic Network on AI Ethics and Policy. International standards bodies were seen as promising for translating red lines into technical and operational requirements.

Rights-focused institutions like the Council of Europe were seen as promising vehicles for championing due process and contestability. And legal bodies such as UNCITRAL were seen as capable of developing model laws and legal templates for domestic adoption, providing a realistic alternative to binding global treaties that may remain out of reach in the current geopolitical climate. Collectively, it was thought that such institutions could help build the standards infrastructure required to turn red lines into verifiable and auditable criteria.

HOW AI RED LINES RELATE TO AI INCIDENTS

Ultimately, the red line domains that surfaced during this year's Athens Roundtable discussions underscore growing global consensus that certain AI use cases, capabilities and behaviors are inherently incompatible with human safety and democratic values. As our Global Call for AI Red Lines campaign aimed to establish, we believe that red lines can serve as critical reference points for establishing a universally shared understanding of what constitutes unacceptable AI threats, and for building common boundaries among different stakeholders around non-negotiable societal protections. They can define the thresholds or conditions under which AI development or deployment becomes intolerable, on grounds that their potential for widespread harm surpasses what any government or organization can responsibly permit.

However, the **effective global governance of severe AI risks and hazards also requires the practical capacity to prevent, detect, and respond to these threats wherever they occur**. **AI incidents and accidents are therefore directly connected to these red lines**. They provide concrete evidence that certain risks have materialized in the real world, rather than being merely hypothetical. When governments observe harmful events or concerning near-misses, this can strengthen the credibility of red line categories, motivate action, and, crucially, inform the establishment of entirely new red lines for previously unconsidered harms that are identified through real-world experience. After all, since the general-purpose nature of and emergent capabilities in powerful AI systems make it impossible to exhaustively anticipate all of the societally unacceptable outcomes they may bring about in advance, incident patterns can reveal where new boundaries must be drawn. Incidents can also function as early warning indicators.

They can reveal when pressure is building towards a particular red line, signaling that unacceptable risks may already be unfolding and thereby creating pressure for governments to monitor, detect, and report such events and to participate in shared incident monitoring and response processes. **Severe AI incidents often transcend borders, so no country can manage them alone**. Cross border monitoring and response infrastructure is therefore essential for collective preparedness, effective prevention, and coordinated action, as it can enable red line conventions to be established, tracked, and enforced globally. This can also support the adoption of robust national and regional frameworks for general-purpose AI governance. If governments both support specific red lines and understand the incident patterns that precede them, they are more likely to put in place the institutional structures necessary to prevent the materialization and escalation of certain kinds of incidents. Such frameworks could thus help countries prevent and manage serious AI incidents and thereby contribute to the stability and security of the global environment.

Taken together, **shared red lines and incident management infrastructure can form a coherent and adaptive AI governance strategy**. **Red lines can set the thresholds** that must never be crossed, while **incident management infrastructure can provide the empirical evidence** needed to refine existing red line thresholds, identify new ones as they emerge, and enable coordinated interventions once thresholds are crossed. Accordingly, focusing on these two areas can enable governments to act early, together, and consistently in the face of severe AI risks.

To find out more about the key takeaways and recommendations that surfaced from the other closed-door dialogue session that took place at this year's Athens Roundtable – focused on the **institutional practices and infrastructure needed to strengthen AI incident prevention and preparedness capacities** – please click here.

ANNEX A - IDENTIFIED RED LINE DOMAINS

Red Line Domain	Unacceptable Risks
CBRN Threats	The use of AI systems by malicious actors for their ability to provide detailed instructions for the development of chemical, biological, radiological, or nuclear weapons.
Loss of Control Risks	AI systems that recursively self-improve (i.e. enhance their own capabilities), autonomously update model weights, or exhibit deceptive scheming capabilities to bypass human oversight.
Cyber Vulnerabilities & Cyber-Offensive Capabilities	The facilitation of large-scale cyberattacks, autonomous zero-day discovery, and the exploitation of security vulnerabilities in critical infrastructure.
Control of Critical Infrastructure	Direct, autonomous control over sovereign compute clusters, power grids, water systems, or financial markets without meaningful human intervention.
Large-Scale Manipulation & Psychological Exploitation	Strategic distortion of human behavior/beliefs (via deception or persuasion) targeting democratic processes; human-AI interactions causing society-wide mental health crises (e.g. psychosis, self-harm, harming others).
Children's Safety	Harms to the development, dignity, and rights of children (e.g. AI-generated child sexual abuse material; stunting of social growth; exploitation of children's data for commercial purposes).
Autonomous Weapons & Military Escalation	The erosion of human control over military engagement and escalation decisions.
Human Rights & Justice	Use of AI to define judicial sanctions, political surveillance, or the absence of a meaningful pathway for redress (contestability).

THE FUTURE SOCIETY

Contact Us

GENERAL: info@thefuturesociety.org
THE ATHENS ROUNDTABLE: aiathens@thefuturesociety.org

www.aiathens.org
www.thefuturesociety.org