THE FUTURE SOCIETY

MARCH 2024

Towards Effective Governance of Foundation Models and Generative AI

TAKEAWAYS FROM THE FIFTH EDITION OF THE ATHENS ROUNDTABLE ON AI AND THE RULE OF LAW

THE
FUTURE
SOCIETY

Towards Effective Governance of Foundation Models and Generative AI: Takeaways from the fifth edition of The Athens Roundtable on AI and the Rule of Law

CONTACT: info@thefuturesociety.org

CITE AS: Amanda Leal. “Towards Effective Governance of Foundation Models and Generative AI: Takeaways from the fifth edition of The Athens Roundtable on AI and the Rule of Law” (The Future Society, March 2024).

THE
FUTURE
SOCIETY

Contents

Contents

Executive Summary

The barriers to a rights-based approach to AI governance anchored in the rule of law have never been more tangible. Increasing AI capabilities, geopolitical tension, and market-driven interests cast doubt on our ability to collectively uphold the public interest in the development and governance of AI systems.

The Athens Roundtable on AI and the Rule of Law is the premier civil society-led multistakeholder forum on AI governance. When the forum was inaugurated in 2019, AI governance frameworks were incipient. The first national strategies were just being developed, international organizations were kickstarting ethical guidelines and seeking stakeholder consensus, and AI policies were far from the spotlight in intergovernmental forums such as the G7 and G20.

Five years later, 2023 marked a year in which AI governance climbed the agenda of policymakers and decision-makers worldwide. The release of technologies with increasingly general capabilities has generated hype and accelerated a concentration of power in big tech, triggering a societal-scale wake-up call. **The growing threats to democratic processes and human rights presented by generative AI systems have prompted calls for regulation. We must collectively demand rigorous standards of safety, security, transparency, and oversight to ensure that these systems are developed and deployed responsibly.**

The fifth edition in numbers

This fifth edition of The Athens Roundtable took place in Washington, D.C., on November 30th and December 1st, 2023. The event brought together **over 1,150 participants in a two-day dialogue focused on coordinating efforts to leverage policy opportunities and co-design actionable solutions**. Discussions focused on governance mechanisms for foundation models and generative AI globally. Participants were encouraged to generate innovative "institutional solutions"—binding regulations, inclusive policy, standards-development processes, and robust enforcement mechanisms—to align the development and deployment of AI systems with the rule of law.

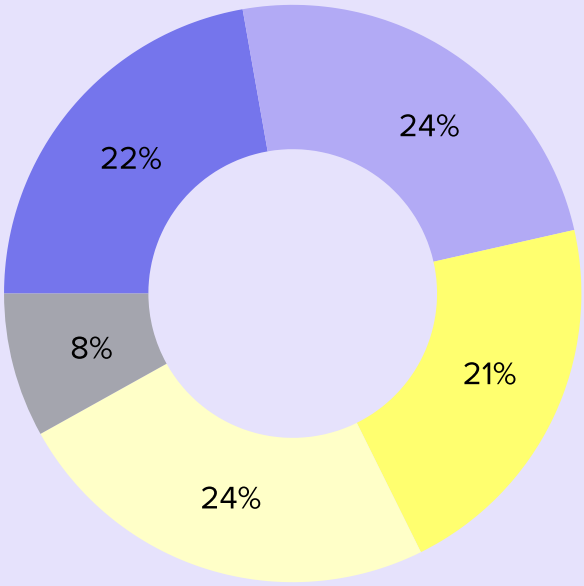
The Roundtable was organized by The Future Society and co-hosted by esteemed partners—the Institute for International Science and Technology Policy (IISTP), the NIST-NSF Institute for Trustworthy AI in Law & Society (TRAILS), UNESCO, OECD, World Bank, IEEE, Homo Digitalis, the Center for AI and Digital Policy (CAIDP), Paul, Weiss LLP, Arnold & Porter, and the Patrick J. McGovern Foundation—and was proudly held under the aegis of the Greek Embassy to the United States. The event welcomed **65 speakers, including policymakers, AI developers, legal experts, and civil society representatives**. Whether on stage, in workshops, in dedicated networking time, or online, the event gathered **an audience of over 200 in-person and 950 online participants, representing over 100 countries in total**. The range of distinguished speakers, including U.S. Senators and Congressmembers, Members of the European and Tanzanian Parliaments, and renowned AI experts, underscored the Roundtable's commitment to a multifaceted and global dialogue on AI governance.

1,150+ ATTENDEES

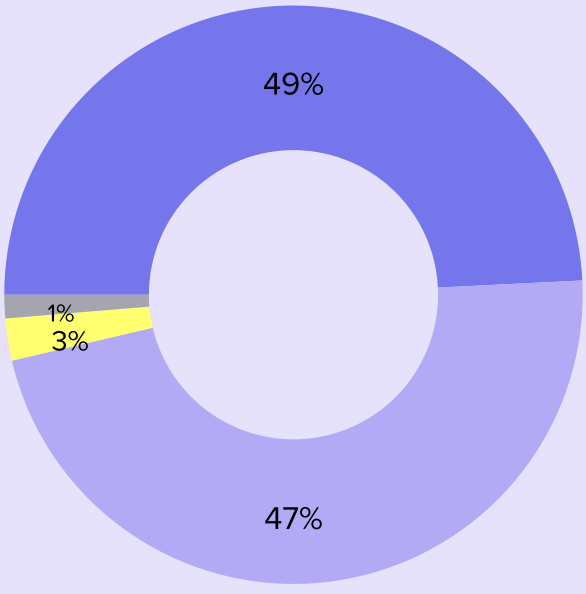
100+ COUNTRIES REPRESENTED

65 SPEAKERS

In-person registrants by sector



In-person registrants by gender



Forward-looking takeaways

The Athens Roundtable informed the public of the latest developments in AI legislation, regulation, standards, and soft governance mechanisms to set appropriate safeguards around foundation models and generative AI. In this context, discussions spanned a broad range of themes, including security vulnerabilities of frontier AI models, policy considerations for open-source AI systems, geopolitical developments, risks of regulatory capture by industry, threats to information ecosystems, and strategies to mitigate the impact of AI on democratic processes.

Discussions probed into national efforts to advance binding regulation, such as U.S. federal legislative efforts, the next steps for federal agencies based on the U.S. Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the European Union's AI Act, and China's generative AI regulation. Dialogues also covered intergovernmental efforts, including the impact of the G7 Hiroshima AI Process on corporate governance, the reach of UNESCO's AI Ethics Recommendation (and subsequent implementation efforts), the potential of a global AI governance initiative stemming from the UN's High-level Advisory Body on AI, and the evidence-based work of the OECD's AI Policy Observatory.

With speakers and participants from over 100 countries, the ideas and arguments presented reflected the viewpoints from a broad range of cultural, political, and socioeconomic backgrounds. **Moving forward, The Athens Roundtable maintains one key commitment: To reexamine our current practices and assumptions, welcoming input and feedback from broad audiences, with particular attention paid to engaging underrepresented communities**.

Below, we present key recommendations that emerged from discussions. These recommendations reflect The Athens Roundtable's mission of advancing responsible AI governance through a harmonized framework encompassing legal compliance and enforcement across jurisdictions. The report that follows presents a session-by-session summary for a more detailed context of discussions.

Key recommendations emerging from discussions

1. ADOPT COMPREHENSIVE HORIZONTAL AND VERTICAL REGULATIONS:
It is crucial that countries adopt legally binding requirements in the form of regulation to effectively shape the behavior of AI developers and deployers towards the public interest. Self- and soft-governance have not realized their promises regarding responsible AI and safety, especially when it comes to foundation models. Sector-specific and umbrella regulations should be adopted in a complementary manner across jurisdictions to fill the existing gap in AI governance. This approach allows for robust governance across the entire AI value chain, from design and development to monitoring, including for general-purpose foundation models that do not fit in any particular sector and may not be covered by current or future sectoral regulations.

REGULATION & ENFORCEMENT

2. STRENGTHEN THE RESILIENCE OF DEMOCRATIC INSTITUTIONS:

There is an urgent need to build resilience in democratic institutions against disruptions from technological developments, notably of advanced general-purpose AI systems. Key elements in building resilience are: capacity-building, in the form of employee training and talent attraction and retention, across government institutions; institutional innovation to bring public sector structures and processes up to date; enforcement authority spanning oversight of the development and deployment of AI systems; and effective public participation. The latter is crucial to ensure that state institutions remain democratic, maintain citizens' trust, and act in the public interest.

TRUST & DEMOCRATIC RESILIENCE

3. ENHANCE COORDINATION AMONG CIVIL SOCIETY ORGANIZATIONS (CSOS) TO ADVANCE RESPONSIBLE AI POLICIES:

In a policy environment with heavy industry lobbying and many conflicting viewpoints, it will be crucial for CSOs to coordinate efforts in order to amplify promising policy recommendations. Key to this coordination will be ensuring that CSOs involved are demographically, culturally, and politically representative of the population at large, and that they consistently listen to the voices of the communities most impacted by emerging technologies.

GLOBAL COORDINATION & CIVIL SOCIETY INCLUSION

4. INVEST IN THE DEVELOPMENT OF METHODS TO MEASURE AND EVALUATE FOUNDATION MODELS' CAPABILITIES, RISKS, AND IMPACTS:

Measurement and evaluation methods play an indispensable role in understanding and monitoring technological capabilities, establishing safeguards to protect fundamental rights, and mitigating large-scale risks to society. However, current methods remain imperfect and will require persistent development in the years to come. Governments should invest in multi-disciplinary efforts to develop measurement and evaluation methods, such as benchmarks, capability evaluations, red-teaming tests, auditing techniques, risk assessments, and impact assessments.

MEASUREMENT & STANDARDS

5. INCLUDE GLOBAL MAJORITY REPRESENTATION AND IMPACTED STAKEHOLDERS IN STANDARD-SETTING INITIATIVES:

Many standard-setting initiatives still lack input from civil society organizations that represent impacted communities. Policymakers and leaders of such initiatives must strive to understand and address structural factors that have led to the under-representation or lack of participation by certain groups in international standard-setting efforts. Potential mechanisms to promote participation include remunerating underrepresented groups and restructuring internal processes to tangibly engage them, rather than provide mere formal representation.

GLOBAL COORDINATION & CIVIL SOCIETY INCLUSION

6. DEVELOP AND ADOPT LIABILITY FRAMEWORKS FOR FOUNDATION MODELS AND GENERATIVE AI:

Liability frameworks must address the complex, evolving AI value chain, so as to disincentivize potentially harmful behavior and mitigate risks. Companies that make foundation models available to downstream deployers across a range of domains benefit from a liability gap, where the causal chain between development choices and any harm caused by the model is currently overlooked. Regulation that establishes liability along the AI value chain is crucial to engender accountability and fairly distribute legal responsibility, avoiding liability being transferred exclusively onto deployers or users of AI systems.

REGULATION & ENFORCEMENT

7. DEVELOP AND IMPLEMENT A SET OF REGULATORY MECHANISMS TO OPERATIONALIZE SAFETY BY DESIGN IN FOUNDATION MODELS:

Given the borderless character of the AI value chain, regulatory mechanisms must be interoperable across jurisdictions. Regulators should invest in regulatory sandbox programs to test and refine foundation models and corresponding regulatory safeguards before deployment.

SAFETY & SECURITY

8. CREATE A SPECIAL GOVERNANCE REGIME FOR DUAL-USE FOUNDATION MODEL RELEASE:

Decisions regarding the release methods for dual-use foundation models should be scrutinized, as they pose societal risks. Exhaustive testing before release would be in the public interest for models at the frontier. Further discussion among stakeholders should identify model release methods that maximize the benefits of open science and innovation without sacrificing public safety.

SAFETY & SECURITY

Moving forward, The Athens Roundtable maintains one key commitment: To reexamine our current practices and assumptions, welcoming input and feedback from broad audiences, with particular attention paid to engaging underrepresented communities.

GLOBAL COORDINATION & CIVIL SOCIETY INCLUSION	**(3)** Enhance coordination among civil society organizations (CSOs) to advance responsible AI policies	**(5)** Include global majority representation and impacted stakeholders in standard-setting initiatives
TRUST & DEMOCRATIC RESILIENCE	**(2)** Strengthen the resilience of democratic institutions	
MEASUREMENT & STANDARDS	**(4)** Invest in the development of methods to measure and evaluate foundation models' capabilities, risks, and impacts	
SAFETY & SECURITY	**(7)** Develop and implement a set of regulatory mechanisms to operationalize safety by design in foundation models	**(8)** Create a special governance regime for dual-use foundation model release
REGULATION & ENFORCEMENT	**(1)** Adopt comprehensive horizontal and vertical regulations	**(6)** Develop and adopt liability frameworks for foundation models and generative AI

Looking ahead, The Future Society remains committed to facilitating dialogues and collaborations. We aim to develop institutional innovations that ensure that the trajectory of AI development aligns with fundamental rights and the rule of law for the benefit of all.

Remarks by The Future Society

Yolanda Lannquist
Director, Global AI Governance
THE FUTURE SOCIETY

REMARKS | Yolanda Lannquist

Yolanda Lannquist emphasized the need for enhanced public and governmental involvement in AI governance. Critiquing the dominance of private sector interests in AI development, Lannquist stressed the need for legislation to implement guardrails that address the safety, security, and ethical risks presented by AI systems. She highlighted **the dangers of prioritizing market growth over safety through the premature launch of advanced AI products**. Lannquist also pointed to some of the risks associated with open access to model weights, such as the ability to remove any existing guardrails. She stressed the importance of establishing proactive policy interventions, monitoring, and accountability mechanisms. Lannquist underscored **the urgency of governing foundation models as they become embedded in consumer applications**.

Nicolas Miailhe
Founder
THE FUTURE SOCIETY

REMARKS | Nicolas Miailhe

Nicolas Miailhe addressed the polarization in AI governance and the need for society as a whole to begin to grapple with the risks associated with AI, from immediate concerns to existential threats. Miailhe stressed the importance of collective action grounded in societal values. Miailhe further pointed to Sam Altman's temporary ouster from OpenAI as underscoring the conflict between public safety and profit-driven motives. To this end, Miailhe suggested, society should demand **legally binding frameworks that prioritize the public interest in AI development**. Miailhe also called for a **tiered governance approach toward AI models based on their capabilities**, a balanced examination of open-source AI's benefits and risks, and **the need to apportion liability appropriately along the AI value chain**, emphasizing the responsibility of lawmakers and policymakers to act.

GLOBAL COORDINATION & CIVIL SOCIETY INCLUSION

KEYNOTE | Vilas Dhar

WATCH NOW

Vilas Dhar | President and Trustee, Patrick J. McGovern Foundation

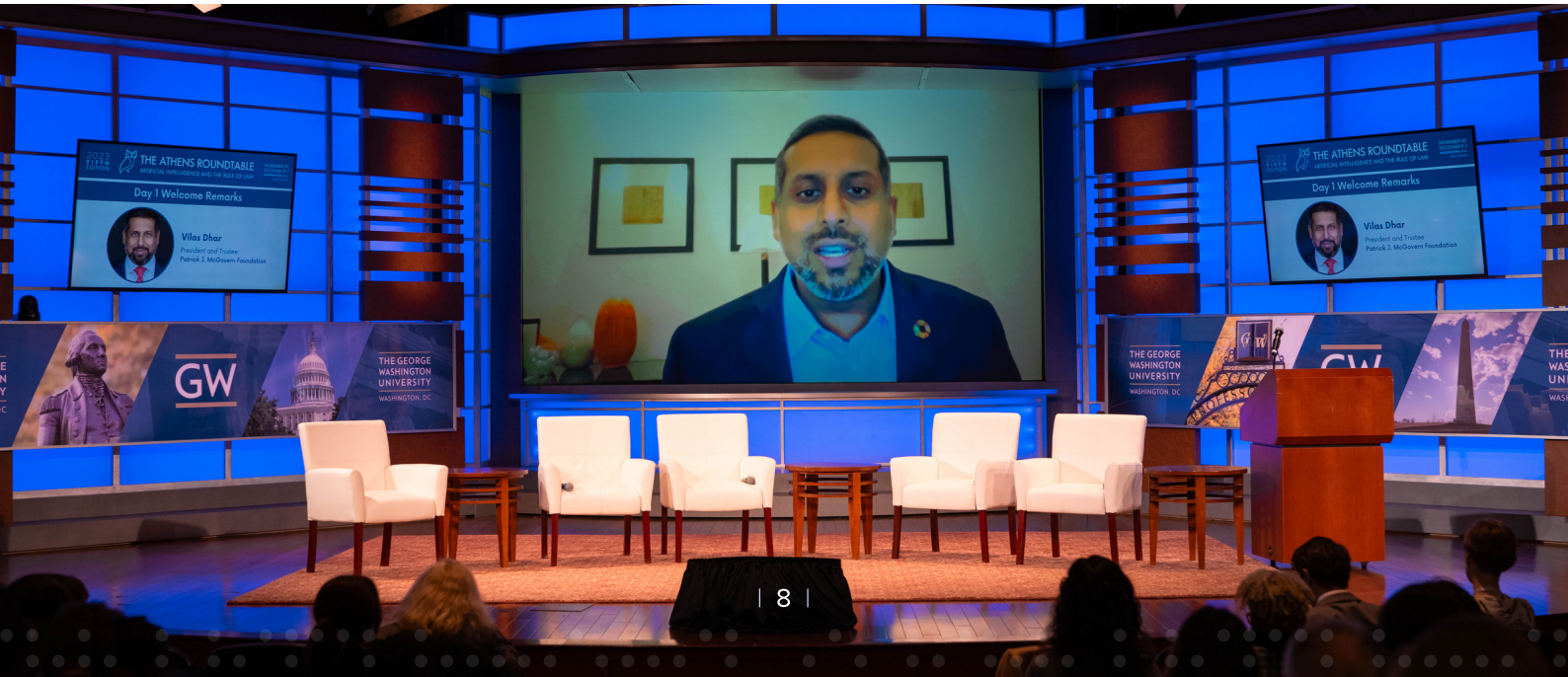
Vilas Dhar, President of the Patrick J. McGovern Foundation, illuminated the deep-rooted, philosophical nature of AI governance discussions, acknowledging the significant contributions of humanists, philosophers, and ethicists over the decades. He stressed the importance of balancing human dignity, justice, and equity with private sector interests, while also creating avenues for solutions that serve universal human interests.

Dhar underscored the critical need for inclusivity in AI policy-making, pointing out **the stark digital divide—with 2.6 billion people still offline—and the dominance of Global North governments in shaping our collective future**. He highlighted the ongoing struggle for civil society, particularly those representing marginalized communities, to gain a meaningful voice in the AI conversation. In the context of the U.S., Dhar criticized the disproportionate focus on big tech narratives and the over-reliance on voluntary self-regulation, which sidelines civil society's participation. Dhar also emphasized the need to include perspectives of the global majority in AI policy, transcending the traditional focus on the US and the EU.

Dhar proposed three key priorities:

1. **Drive public investment to bridge the digital divide** and ensure broadband connectivity for all.
2. **Address the data gap in AI development** by prioritizing the collection and use of diverse data sets that truly represent global populations, especially in areas like health and drug discovery.
3. **Leverage policy mechanisms to close the technical capacity gap among the global majority**. This is paramount to fostering a diverse community of socially conscious AI practitioners.

Dhar concluded with optimism, highlighting the potential for shared values to lead to policy harmonization, ultimately benefiting economic, social, and political outcomes around the world.



PANEL | Global Governance: What's next for international institutions

WATCH NOW

Amandeep Singh Gill | Secretary Generals' Envoy for Technology, United Nations
Gabriela Ramos | Assistant-Director General for Social and Human Sciences, UNESCO
Ulrik Vestergaard Knudsen | Deputy Secretary-General, OECD
Gary Marcus | Emeritus Professor of Psychology and Neural Science at NYU and CEO at the Center for the Advancement of Trustworthy AI
Susan Ariel Aaronson (moderator) | Professor of Intl. Affairs, Director of the Digital Trade and Data Governance Hub, and co-PI of NIST-NSF TRAILS, The George Washington University

Main Takeaways

ENSURE THE RESILIENCE OF DEMOCRATIC INSTITUTIONS WITH REGULATION AND CAPACITY-BUILDING:
Given the host of unintended outcomes they may occur through the development and deployment of large models, it is urgent to strengthen democratic institutions and develop mechanisms to mitigate harms when they happen. This should be done through increased capacity-building in the public sector and a move towards binding laws.

IMPLEMENT GLOBAL PRE-DEPLOYMENT SAFETY MECHANISMS:
To mitigate the potential widespread harms of foundation models, speakers recommended international coordination toward developing and deploying ex-ante evaluations, risk assessment methodologies, human rights impact assessments, red teaming, AI capability control mechanisms, interoperable AI auditing practices, and certification ecosystems.

SHAPE AI GOVERNANCE FOR THE PUBLIC INTEREST:
The panel highlighted the urgency to think beyond safety and harm towards a broader notion of public interest, including AI's impact on human rights, environmental sustainability, and economic inclusion. Speakers expressed concern about the growing concentration of power in AI, and stressed the importance of the involvement of a diverse range of stakeholders in AI governance initiatives.

Discussion

This panel addressed two questions that rose to the core of international AI governance discussions in 2023. First, how can we collectively facilitate an international AI governance regime? Second, what safety mechanisms are needed to address the borderless impact of foundation models?

OECD's Deputy Secretary-General Ulrik Vestergaard Knudsen opened the discussion by classifying the rise of generative AI as a watershed moment, calling for updates on international institutions' work. He detailed the **OECD's plans to review its 2019 principles in light of evolving AI capabilities** and the OECD's ongoing efforts to collect evidence of AI impacts and strengthen its AI experts community. Knudsen acknowledged the OECD's inherent focus on only a subset of countries, but called for broader collaboration in AI governance to facilitate the convergence of approaches.

UN Tech Envoy Amandeep Gill stressed the need for an **inclusive network of institutionalized responses to AI**, in which different international institutions and blocks of countries share knowledge and coordinate harmonized approaches. He highlighted the UN Tech Envoy's mandate, whose advisory body is developing a **comprehensive risk assessment framework encompassing short- to long-term risks**. Dr. Gill is also focusing on democratizing AI opportunities aligned with sustainable development goals (SDGs). Key challenges include updating practices and responses promptly in a fast-evolving AI landscape, improving multistakeholder participation, and coordinating responses across industry, civil society, and governments at the national and international levels.

Adding to the multistakeholder challenge, he noted the **looming risk of regulatory capture, given the high concentration of power in a few corporations**. Finally, Dr. Gill emphasized that transparency in data source disclosure remains an underexplored area in AI governance with global implications.

UNESCO's Assistant Director-General for Social and Human Sciences, Gabriela Ramos, underlined the necessity of **ex-ante AI assessments** and adherence to human rights standards. She described **UNESCO's collaboration in creating an AI ethics observatory** with civil society organizations to ramp up analytical efforts to inform policymaking. In addition to analytical work, UNESCO has been actively working with national governments to apply readiness assessments and build public sector capacity to implement the Recommendation on the Ethics of Artificial Intelligence, adopted by 194 countries. Reflecting on growing concerns about generative AI and foundation models, Ramos stressed the importance of **legal responsibility and liability frameworks for AI developers**, highlighting the challenge of implementing these globally.

Gary Marcus dissected the shift in the public discourse from "trust" to the "safety" of AI systems, pointing out how events such as ChatGPT's launch contributed to growing safety-oriented concerns. He criticized the **rapid commercialization of AI before robust safety guardrails were in place**, exemplified by various instances of dangerous outputs by AI systems, as was the case with Microsoft's Sydney. Finally, Dr. Marcus called for work to **increase reproducibility in the development of AI systems**.

Reflecting on growing concerns about generative AI and foundation models, Ramos stressed the importance of legal responsibility and liability frameworks for AI developers...

FIRESIDE CHAT | **Trustworthy AI in the Global South**

WATCH NOW

Neema Lugangira | Member of Parliament Tanzania, and Chair, African Parliamentary Network on Internet Governance
Yolanda Botti-Lodovico (moderator) | Policy and Advocacy Lead, Patrick J. McGovern Foundation

Main Takeaways

LAWMAKERS HAVE A KEY ROLE IN INTERNATIONAL LEGAL AND REGULATORY HARMONIZATION:
Lawmakers can advance AI governance coordination in their respective jurisdictions by enacting laws, voting on and proposing investment priorities, and raising awareness within government, among colleagues, and with constituents. Lawmakers, as a crucial stakeholder group, must be engaged in discussions and convenings on AI governance.

DEVELOP LIABILITY FRAMEWORKS AND REGULATIONS THAT ADDRESS THE GLOBAL CHARACTER OF THE AI SUPPLY CHAIN AND ENABLE REDRESS FROM IMPACTED PEOPLE:
Lawmakers, especially those of advanced economies, must not overlook the impact on the Global South through companies' supply chains. Countries should enact liability frameworks that hold actors accountable and strengthen corporate responsibility beyond their borders.

EXPAND REPRESENTATION AND PARTICIPATION OF THE GLOBAL MAJORITY IN INTERNATIONAL AI GOVERNANCE DECISION-MAKING PROCESSES:
Recent high-level convenings on AI governance have had limited geographic representation, which puts at risk any outcome with global ambitions. Global AI governance decision-making processes should strive to include global majority representatives across different stakeholder groups: governments, independent academic experts, civil society representatives, and industry representatives.

Discussion

In an insightful conversation on trustworthy AI, Honorable Member of the Tanzanian Parliament Neema Lugangira and policy expert Yolanda Botto-Ludovico highlighted the necessity for global inclusivity in decision-making, knowledge sharing, and capacity building, as well as robust regulatory frameworks that prevent exploitative dynamics in the Global South. The discussion underscored the importance of **equitable AI development, international collaboration, and accountability in AI governance**, emphasizing that the benefits of AI should be democratized globally in a secure, safe, and ethical manner.

Drawing from her experience in Tanzania and internationally in the Inter-Parliamentary Union (IPU) and the African Parliamentary Network on Internet Governance, MP Lugangira pointed out the crucial need to equip lawmakers with the knowledge and tools to contribute to AI governance effectively. She advocated for increased participation of parliamentarians in AI governance discussions worldwide and underscored the urgent need for global legislative attention on AI's societal implications. In her role at the Inter-Parliamentary Union, **MP Lugangira is co-sponsoring a draft resolution for the IPU's first general assembly of 2024**, focusing on the impact of AI on democracy and the rule of law. If approved, the resolution can influence national legislative discussions across the 193 countries represented at the IPU.

While **Global South countries are economically, politically, and socially affected by the development of AI systems, they have often not been meaningfully represented in international AI governance convenings** such as the UK AI Safety Summit. MP Lugangira emphasized the urgency of transforming African countries from mere consumer markets to respected and active participants and developers in the AI landscape. A positive step toward that direction is the African Union's ongoing collaboration with the OECD AI in writing a continental AI Strategy. This strategy will be key in leveraging AI to address critical challenges like food insecurity on the African continent.

Drawing attention to the global majority's crucial role in AI development, MP Lugangira highlighted that **data is the backbone of foundation models**. She raised concerns about companies exploiting data from African nations without compensation and stressed the importance of enacting **laws across jurisdictions allocating liability throughout the global AI supply chain**. Countries in the Global North developing AI regulations should set a standard of behavior that allows individuals in the Global South to hold AI companies accountable for harms reproduced beyond companies' headquarters jurisdictions.

Hon. MP Lugangira pointed out the crucial need to equip lawmakers with the knowledge and tools to contribute to AI governance effectively. She advocated for increased participation of parliamentarians in AI governance discussions worldwide and underscored the urgent need for global legislative attention on AI's societal implications

FIRESIDE CHAT | The Path to Generative AI Regulation in China

WATCH NOW

Yi Zeng | Professor, Director; Brain-inspired Cognitive Intelligence Lab and International Research Center for AI Ethics and Governance, Institute of Automation, Chinese Academy of Sciences; Founding Director, AI for SDGs Cooperation Network; Founding Director, Center for Longterm AI
Samuel Curtis (moderator) | Senior Associate, The Future Society

Main Takeaways

EMBRACE A COMPLEMENTARY REGULATORY APPROACH:
Countries should learn from each other, with a focus on integrating both vertical approaches, as China has applied in some sectors, and horizontal approaches, akin to the EU's for comprehensive AI governance.

INCREASE THE PARTICIPATION OF INDEPENDENT ACADEMIC EXPERTS:
Prof. Zeng advocated for increasing academic experts' participation in high-level advisory groups, like the United Nations High-Level Advisory Body on AI, to ensure a diversity of perspectives and a balanced approach to AI governance.

BROADEN THE INTERNATIONAL DIALOGUE TO INCLUDE DIVERGENT VIEWPOINTS:
Global coordination efforts must move beyond discussions among "like-minded" countries to more inclusive and substance-oriented dialogues. It is urgent to bridge diverse perspectives and foster a collaborative international environment for AI development and regulation.

Discussion

In a conversation with Professor Yi Zeng, a renowned expert in AI ethics and governance, TFS's Samuel Curtis inquired about Prof. Zeng's views on China's unique approach to AI governance and its contributions to the global AI regulatory landscape.

During the discussion, Yi Zeng emphasized **the symbiotic nature of various regulatory approaches**, contrasting the European Union's broad, horizontal AI Act with some of China's more targeted, vertical regulations focusing on specific AI applications, such as recommendation systems and generative AI. Prof. Zeng highlighted the benefits of China's approach, particularly its specificity in addressing AI challenges. He also suggested that, conversely, China could learn from the EU's broader, horizontal framework.

In his analysis of international AI safety commitments —including the UK AI Safety Summit, the Bletchley Declaration, and the joint statement signed by world-leading academics highlighting the importance of AI safety, called the Ditchley Declaration—Prof. Zeng emphasized the necessity of inclusive dialogue in global coordination efforts. He stressed the importance of achieving a unified, global understanding of the risks presented by AI development and that this will require engaging with nations across the geopolitical spectrum. The participation of China in the summit was an example of this inclusive approach, ensuring that discussions on AI safety encompass a diversity of cultural and political viewpoints.

In addition to these policy discussions, Prof. Zeng underscored the crucial role of academia in shaping AI governance. He highlighted the unique contributions of independent academic experts in providing balanced, interdisciplinary perspectives on AI's long-term risks. Prof. Zeng advocated for **independent academic expertise to guide AI development beyond national competition,** focusing on collaborative problem-solving.

Prof. Zeng emphasized the necessity of inclusive dialogue in global coordination efforts. He stressed the importance of achieving a unified, global understanding of the risks presented by AI development and that this will require engaging with nations across the geopolitical spectrum.

REMARKS | **Yoichi Iida**

WATCH NOW

Yoichi Iida | Deputy Director General, Ministry of International Affairs and Communication, Japan

Yoichi Iida joined The Athens Roundtable to celebrate the completion of the first phase of the Hiroshima AI Process, with the guiding principles and the **G7 code of conduct**. He shared a reflection on his experience as a representative of Japan and chair of the **Hiroshima AI Process Working Group**, which took place on December 1st, 2023.

Mr. Iida commented on the Hiroshima AI Process report, which includes a **comprehensive framework that promotes safe, secure, and trustworthy generative AI**. Notably, it was developed with the intention of being adopted beyond G7 member countries. The framework is comprised of four elements: The OECD report covering the potential risks, challenges, and opportunities brought by generative AI and foundation models; guiding principles for AI actors across the value chain; a set of measures and actions targeted at advanced AI (generative AI and foundation models) developers, including the code of conduct; and a set of projects that will explore potential solutions to respond to emerging risks and challenges of those technologies. The projects aim to tackle **foundation models' lack of transparency and the spread of AI-enabled disinformation**.

Mr. Iida commended the Hiroshima AI Process Working Group's strong commitment to collaborating with a variety of stakeholders across the world and other multilateral organizations to operationalize the framework.

> *... the Hiroshima AI Process report ... was developed with the intention to be adopted beyond G7 member countries. The framework is comprised of four elements: The OECD report covering the potential risks, challenges, and opportunities brought by generative AI and foundation models; guiding principles for AI actors across the value chain; a set of measures and actions targeted at advanced AI developers, including the code of conduct; and a set of projects that will explore potential solutions to respond to emerging risks and challenges of those technologies.*

REMARKS | **U.S. Representative Sara Jacobs**

WATCH NOW

Sara Jacobs | United States Representative

U.S. Representative Jacobs highlighted the need for globally coordinated AI governance, stressing the need to address AI's impact on the Global South, and advocated for inclusive, multilateral engagements.

U.S. Representative Jacobs underscored the importance of incorporating diverse voices, especially Civil Society Organizations (CSOs) and governments from the Global South, into AI policy discussions. She also urged for a comprehensive approach to AI safety, **encouraging the AI governance community to develop a broad definition of "AI safety," encompassing the entire spectrum of AI risks**. It is particularly important to address bias and ongoing harms incurred by marginalized communities for increased safety.

Finally, U.S. Representative Jacobs echoed calls for robust oversight and regulation. She encouraged leading nations, especially the US, to adopt new legislation and develop new resources for AI oversight. She stressed the strategic role new institutions like the **U.S. AI Safety Institute** can play in fostering adaptive and effective AI governance.

TRUST & DEMOCRATIC RESILIENCE

PANEL | The impact of Generative AI on Elections

WATCH NOW

Dr. Rebekah Tromble | Director, Institute for Data, Democracy & Politics, George Washington University
Caio Machado | Executive Director, Instituto Vero
Marielza Oliveira | Director of the UNESCO Communications and Information Sector's Division for Digital Inclusion, Policies, and Transformation
Paul Nemitz | Principal Adviser on the Digital Transition, European Commission
Merve Hickok (moderator) | President, Center for AI and Digital Policy

Main Takeaways

REGULATE THE FINANCING OF DIGITAL CAMPAIGNS AND THE USE OF MICROTARGETING FOR ELECTORAL PURPOSES:
The unregulated use of generative AI in electoral campaigns is extremely harmful to democracy. Governments must ensure that electoral outcomes don't hinge more on financial resources and technical capabilities than on democratic discourse and voter engagement.

ENSURE THE SAFETY OF JOURNALISTS AND PROTECT AUTHENTIC CONTENT:
Emphasizing the need to protect professional and evidence-based journalism, the panelists called for stringent redressing measures against attacks on journalists. Policymakers must consider coordinating toward global norms to manage the influx of AI-generated content and misinformation. Other urgent measures to preserve the digital ecosystem's integrity include enforcing standardized guidelines for platform governance, providing incentives for authentic content creation, and mechanisms to label AI-generated or AI-altered content, such as watermarks.

ESTABLISH AI GOVERNANCE EXPERT GROUPS FOR ELECTIONS:
Speakers proposed the creation of a group that provides AI governance support at scale and on-demand, especially for electoral bodies in regions with fragile electoral systems. Independent expert groups can help enhance the integrity and security of electoral processes worldwide.

Discussion

In 2024, approximately 3.9 billion people—48% of the world's population—will participate in general elections across 54 countries, including five nuclear-armed states. At this critical political juncture, policies must address AI's potential to amplify disinformation, heighten cybersecurity threats, and disrupt the information ecosystem.

Opening the discussion, Dr. Rebekah Tromble analyzed the potential impact of generative AI on voter behavior and election dynamics. She emphasized the crucial role of AI transparency in the upcoming 2024 U.S. presidential elections. Dr. Tromble advocated for **data and model disclosure requirements** and emphasized the need to educate and inform the public about the potential impact of AI on voter behavior and election dynamics.

The discussion followed with an analysis of the adequacy of the current European regulatory landscape in addressing the challenges posed by generative AI and disinformation. Paul Nemitz defended **the vetting of large AI models by regulators before public release,** similar to the safety requirements of other industries such as automotive and pharmaceuticals. Focusing on elections, Nemitz emphasized the dangers of electoral outcomes' greater dependence on digital advertising and sophisticated targeting technologies than on the quality of political arguments or candidates' trustworthiness. Nemitz stressed the urgent need to **rethink current models of election advertisements and party financing**. Reflecting on the Digital Services Act and the European context, Nemitz highlighted that jurisdictions that adopt platform and electoral advertising regulations with policies to combat disinformation would be less likely to have elections swayed by digital targeting techniques.

Caio Machado criticized the resistance tech companies often exhibit towards regulation. **The absence of institutional mechanisms to diagnose problems and understand the impact of technologies creates a trust gap, which may lead to radical and hasty, non-technical solutions or simply inaction.** He also pointed out the need to rethink information production, usage, validation, and dissemination. Machado shared his experience combating disinformation in Brazil, where Institute Vero collaborated with social media creators to educate young people and judiciary staff on fact-checking and open-source investigation tools. Regulations and institutional mechanisms should be tailored to their respective local contexts while pursuing a global goal: **strengthening democratic institutions' resilience to technological disruption**.

In her response, Marielza Oliveira from UNESCO's Division for Digital Inclusion, Policies and Transformation articulated the organization's approach to balancing freedom of expression with

Regulations and institutional mechanisms should be tailored to their respective local contexts while pursuing a global goal: strengthening democratic institutions' resilience to technological disruption.

oversight for a democratic information ecosystem, rooted in Article 19 of the International Covenant for Civil and Political Rights. She underscored **the threats of the rapid diffusion of harmful AI-generated content on social media platforms. In the Global South, infrastructure and skills gaps exacerbate disparities in the information ecosystem.** To fight the surge of AI-generated content, Dr. Oliveira emphasized UNESCO's efforts in assessing digital ecosystems: Currently, 44 countries are voluntarily assessing the extent to which their digital ecosystems are human rights-based, open, accessible, and multi-stakeholder-led. These assessments help identify and address systemic shortcomings, such as issues related to language barriers, accessibility, and inclusion.

Speakers analyzed possible restrictive measures to protect electoral integrity. The discussion also touched upon the need to rethink business models that significantly impact democracy and the rule of law. Micro-targeting, for instance, is widely accepted in the economic realm but problematic for electoral integrity.

Furthermore, society must be equipped with critical thinking skills, media literacy, and access to information to properly make sense of the digital ecosystem. Society must demand accountability of tech corporations, politicians, and governments for the development and deployment of potentially harmful AI systems. Speakers also emphasized the importance of supporting trustworthy information sources, journalists, and local news in light of AI-generated content.

Finally, speakers concluded that transparency alone is insufficient for preserving democracy. **Governments should adopt positive agendas focused on rebuilding trust,** lest environments of systemic distrust undermine the political institutions themselves.

FIRESIDE CHAT | **Uncovering the Use of Generative AI in Judicial Contexts**

WATCH NOW

Juan David Gutierrez Rodriguez | Associate Professor, School of Government of Universidad de los Andes
Miriam Stankovich | Principal Digital Policy Specialist, DAI
Linda Bonyo | Founding Director, Africa Law Tech; Founder, Lawyers Hub Kenya
Kimberly H. Kim | Assistant Chief Administrative Law Judge, California Public Utilities Commission
Cédric Wachholz (moderator) | Chief of Section, Digital Innovation and Transformation Section, UNESCO

Main Takeaways

DEVELOP AND DEPLOY FORMAL TRAINING AND GUIDELINES FOR THE USE OF AI IN THE JUDICIARY:
The panel underscored the need for formal training and guidelines on the judicial use of AI, emphasizing that AI tools must be leveraged in an informed, ethical, and responsible manner. UNESCO has been leading efforts in this area, with the MOOC on AI and the Rule of Law co-produced with The Future Society, the Toolkit on AI and the Rule of Law, and the upcoming guidelines for the use of generative AI in judicial contexts. Acknowledging the diverse impacts of AI across different regions, the discussion highlighted the need for contextual adaptations of training and guidelines and equitable access to resources, particularly in the Global South.

IMPLEMENT INSTITUTIONAL INNOVATION IN THE JUDICIARY TO LEVERAGE AI TO EXPAND ACCESS TO JUSTICE:
The conversation pointed towards future-proofing the judiciary against the challenges posed by AI, including developing adequate mechanisms to manage increased caseloads and reviewing processes to ensure fair access to justice in an increasingly digital legal landscape.

Discussion

This fireside chat brought together experts from different corners of the world to discuss the burgeoning intersection of generative AI and the judiciary. Moderated by Cédric Wachholz, panelists shed light on AI utilization in judicial settings, the associated risks, and the pressing need for guidelines and training in this domain.

Professor Juan David Gutierrez Rodriguez initiated the conversation by sharing findings from UNESCO's global survey on generative AI in the judiciary. **This survey, receiving responses from nearly 100 countries, indicated a significant gap between legal operators' familiarity with AI and its practical application in professional legal contexts.** While most respondents were acquainted with AI, only a fraction used AI tools, such as large language models, for professional purposes. Concerns that were highlighted included data security, reliability of information, and potential violations of privacy and copyright. Notably, a vast majority of the respondents had not received formal AI training, indicating a dire need for educational initiatives and mandatory rules to govern the use of AI in legal settings.

Miriam Stankovich underscored the challenges posed by generative AI tools in the judiciary. Drawing from her collaboration with UNESCO on the recently launched Global Toolkit on AI and the Rule of Law, she pointed out the tendency of generative AI tools to create plausible but not necessarily accurate outputs, "hallucinate," and amplify bias. Stankovich emphasized **the urgent need for enhanced regulation, governance, and, critically, digital literacy among judges to navigate the complexities of AI in judicial proceedings.**

Linda Bonyo focused on the context-specific impact of technology, highlighting the **disparities in AI tool performance and adoption across different regions**, particularly in Kenya and the broader Global South. She pointed out that when procuring cutting-edge technologies, such governments often must rely upon products developed in the Global North, emphasizing the issue of vendor lock-in due to the lack of viable local alternatives. Bonyo further advocated for **equitable access to computing resources in the Global South,** helping countries transcend the roles they might otherwise be confined to—mere consumption and data labeling.

Judge Kimberly Kim provided a cautiously pragmatic perspective on the use of generative AI in the U.S. judiciary. She highlighted the judiciary's resistance to technological changes and the need for tools and programs to educate legal operators about generative AI. Preparing the judiciary for the operational impacts of AI is urgent. Courts might see an increase in dockets due to AI-related claims and lawsuits. She furthermore stressed the urgent need for equitable access to justice: **costly AI-assisted legal tools will confer strategic legal advantages** to those with access to them, which will likely **contribute to a growing digital divide in the justice system.**

Professor Rodriguez concluded the panel with insights into upcoming UNESCO recommendations on generative AI use in judicial contexts. He emphasized **the need to test AI tools before deployment, particularly in high-risk environments like the justice sector**, and to assess their impact on human rights.

FIRESIDE CHAT | **Trends and Challenges for AI Governance in 2024**

WATCH NOW

Keith Sonderling | Commissioner, U.S. Equal Employment Opportunity Commission (EEOC)
Peter Schildkraut | Technology, Media & Telecommunications Industry Team Co-Leader, Arnold & Porter
Tawana Petty | 2023-2025 Just Tech Fellow, Social Science Research Council
Nicolas Moës (moderator) | Executive Director, The Future Society

Main Takeaways

DEVELOP STRATEGIES AND ORGANIZATIONAL STRUCTURES THAT ALLOW FOR CIVIL SOCIETY ORGANIZATIONS TO UNITE IN ADVOCATING FOR PROMISING AI POLICIES:
Civil society is composed of actors with a wide range of priorities, values, and backgrounds, but they share a common goal of advancing the public interest. Civil society organizations should collectively identify and advance promising AI policies to counteract growing corporate lobbying efforts.

STRIVE FOR DIVERSITY IN CORPORATE COMPLIANCE AND RISK MANAGEMENT BOARDS:
Internal boards should be capable of foreseeing a range of risks and recommending appropriate actions. A greater diversity of perspectives would enable companies to identify different layers of risks and foresee negative outcomes, from product development to post-deployment maintenance.

FACILITATE THE PARTICIPATION OF IMPACTED COMMUNITIES IN REGULATORY PROCESSES:
Tawana Petty highlighted the power of grassroots movements and local organizations in shaping AI policy, stressing the importance of including voices often sidelined in the regulatory process. In establishing appropriate safeguards, regulatory discussions must consult with impacted communities and consider case studies that elucidate the impact of emerging technologies on the ground.

PRIORITIZE THE EXECUTIVE POWER'S AI CAPACITY-BUILDING:
Government agencies must urgently develop expertise and capabilities to understand, audit, and effectively regulate AI technologies. Agencies should seek knowledge from independent academic experts to fill the current skills gap.

Discussion

This panel centered on the evolving landscape of AI governance in 2024, exploring the balance between technological innovation and the need for robust regulatory frameworks. The conversation highlighted the importance of diverse stakeholder involvement in shaping AI policy and the challenges of adapting existing regulatory mechanisms to the nuanced demands of AI technologies.

Keith Sonderling emphasized **the significance of involving new stakeholders, such as auditors and AI developers, in the regulatory space, particularly in the context of employment and civil rights.** He highlighted the necessity of integrating these new perspectives to ensure AI technologies are developed and deployed without discriminating against marginalized groups. Government agencies such as the EEOC have a crucial role to play, regardless of legislative developments and new regulations, in enforcing existing laws in cases pertaining to the use of AI and its impact on people, as well as applying rigorous oversight to the deployment of AI in the sectors they are mandated for. The use of natural language processing tools in hiring assessments, for instance, might discriminate against non-native English speakers or people with speech impairments, which falls under the purview of the EEOC.

Peter Schildkraut discussed the judiciary's vital role, especially in the U.S., in defining rights and addressing AI-related harms. He provided insights into ongoing litigation that illustrates the judiciary's role in clarifying the application of existing laws to AI technologies, thus shaping the trajectory of AI regulation. In addition, Schildkraut analyzed the industry's emerging challenges for compliance and risk management. Notably, **companies should ensure their risk- and impact-assessment bodies are comprised of professionals with technical AI expertise and professionals with different backgrounds and subject matter expertise.** Finally, they should be empowered within the company to make decisions to discontinue the development or production of dangerous AI models.

Centering the main challenge for 2024 on AI policy and representation, Tawana Petty stressed that, although the US has advanced in developing AI governance frameworks such as the Blueprint for an AI Bill of Rights, **there are still missing voices in the dialogues and decision-making processes pertaining to AI governance.** Petty underscored the potential of people and grassroots movements in influencing policy and demanding inclusion when civil society groups are marginalized in decision-making processes. She advocated for the **inclusion of diverse voices—particularly those most impacted by AI technologies—in regulatory discussions.** In order to advance responsible AI policy, civil society organizations should coordinate efforts, leveraging intersectionality and elevating marginalized voices rather than seeking a monolithic approach.

In order to advance responsible AI policy, civil society organizations should coordinate efforts, leveraging intersectionality and elevating marginalized voices rather than seeking a monolithic approach.

REMARKS | **Cédric Wachholz**

WATCH NOW

Cédric Wachholz | Chief of Section, Digital Innovation and Transformation Section, UNESCO

Cédric Wachholz shared insights from a recent survey conducted among UNESCO's network of 35,000 judicial operators from over 100 countries, focusing on generative AI and its role in the judiciary. The survey revealed **a dramatic increase in AI use within legal systems worldwide, raising important questions about AI's role in enhancing justice while upholding human rights and democratic values.**

While there had initially been an international convergence towards a multidimensional approach to AI governance focused on the protection of human rights, recent industry developments suggested market pressures often prioritize profit over safety and private interests over public ethical AI governance.

Meanwhile, governments are grappling with fostering innovation-friendly environments while establishing clear, effective AI guardrails. In the judiciary, AI offers potential benefits in decision-making, access to justice, and crime prevention.

However, cases in which AI systems are the object of litigation remain markedly complex. Wachholz cited a case in Brazil where the use of "smart billboards" in the São Paulo metro system to predict riders' emotions and other attributes was challenged.

Wachholz also mentioned **UNESCO's significant role in training judicial operators and the growing demand for AI training.** He referenced the AI and the Rule of Law MOOC, launched by UNESCO in partnership with The Future Society and other organizations, which has educated over 5,900 judicial operators from 141 countries. Additionally, UNESCO recently introduced a Global Toolkit on AI and the Rule of Law for the judiciary.

In conclusion, Wachholz called for collaborative efforts to transform discussions into action, urging participants to work together to ensure AI supports rather than undermines justice.

SAFETY & SECURITY

KEYNOTE | Yoshua Bengio

WATCH NOW

Yoshua Bengio | Scientific Director, Mila & IVADO, Full Professor, Samsung AI Professor, Université de Montréal, Canada CIFAR AI Chair

In an inspiring keynote, Professor Yoshua Bengio laid out his perspective on countering safety and security risks of increasingly capable AI systems, examining how the balance of power is pivotal for the survival of democracies.

Prof. Bengio identified **two primary technical challenges confronting AI today: the risks to security and the looming threat of losing control over AI systems.** He illuminated the difficulties in training AI systems that are assuredly safe and the ease with which malign actors could exploit open-source AI systems. Furthermore, he discussed the scientific community's debate over AI systems potentially developing self-serving objectives, deviating from human interests. Prof. Bengio pointed out the current scientific limitations in ensuring that AI systems align with human intentions and interests, which is evident in existing biases and discrimination in AI systems.

Evaluating President's Biden *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Prof. Bengio commended it as a pivotal step towards bolstering AI governance. The order's approach to measuring and evaluating AI systems based on their computational resources used in training was highlighted as a key development. He commented that presently, compute utilization is a reasonable proxy for models' capability. The more capability, the more potential to create harm. However, **we must continue to develop more robust measurement and evaluation mechanisms**.

Prof. Bengio also emphasized the responsibilities of companies and governments in the AI domain: **Companies should proactively demonstrate the safety of their AI systems, and governments should develop capacity in AI measurement and evaluation for effective oversight.**

On regulatory measures, Prof. Bengio advocated for strategies to mitigate risks associated with AI systems falling into the wrong hands. He suggested implementing **licensing regimes, reporting requirements, and auditing for the most powerful AI systems**. He stressed the need for greater scrutiny before highly capable models are released open source. Vulnerabilities in open-source models can't be retroactively fixed throughout the value chain once models have been downloaded. Prof. Bengio underscored that decisions on releasing these models should involve a democratic evaluation process due to the significant global risks involved.

Finally, Prof. Bengio called for **institutional innovation in democratic processes to control AI development**. He proposed the formation of multi-stakeholder governance bodies, comprising civil society, academics, and media, to oversee AI development and ensure societal alignment.

> *Prof. Bengio identified two primary technical challenges confronting AI today: the risks to security and the looming threat of losing control over AI systems.*

KEYNOTE | U.S. Representative Anna Eshoo

WATCH NOW

Anna Eshoo | United States Representative

U.S. Representative Anna Eshoo, in her keynote, addressed the duality of AI as a source of both groundbreaking advancements and potential perils. She emphasized the need for AI development to be safe, trustworthy, and responsible, highlighting the importance of these qualities in the context of rapid technological progress.

Representative Eshoo outlined three fundamental requirements for AI research and development: access to good data, sufficient computing power, and skilled people. She argued against the monopolization of AI development by large technology companies, advocating for a more inclusive approach. She emphasized that startups, small businesses, academia, medical and non-profit communities, and the public sector should all have access to essential AI resources.

Representative Eshoo stressed that **democratizing AI research and development would enable researchers and innovators across the United States to develop AI tools that bolster our national security, advance safety and economic competitiveness, and improve society in numerous ways**. To this end, she deemed it critical that the U.S. Congress passes the **CREATE AI Act**, a bipartisan and bicameral piece of legislation that would establish the national AI research resource, a shared cyber research infrastructure.

The convergence of biosecurity and AI is another area of concern demanding regulation, stressed Representative Eshoo. President Biden's *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* directed agencies to conduct a study on how AI can increase biosecurity risks, aligned with concerns brought to Congress by Representative Eshoo, such as the need for AI and biosecurity risk assessment.

Finally, Representative Eshoo underscored national security as Congress's top priority. As **AI doesn't recognize any national boundaries, it is also imperative to work on international coordination** to advance AI governance that reflects fundamental values, protects our democracy, and respects the rule of law.

> *Representative Eshoo ... argued against the monopolization of AI development by large technology companies, advocating for a more inclusive approach. She emphasized that startups, small businesses, academia, medical and non-profit communities, and the public sector should all have access to essential AI resources.*

PANEL | Managing Safety and Security of Foundation Models

WATCH NOW

Irene Solaiman | Head of Global Policy, Hugging Face
Joslyn Barnhart | Senior Research Scientist, Strategic Governance Lead, Google DeepMind
Tom Goldstein | Volpi-Cupal Professor of Computer Science, University of Maryland; Co-PI, NIST-NSF TRAILS
Juraj Čorba | Digital Regulation & Governance Expert, Slovak Ministry of Investments, Regional Development and Informatization; Chair-Elect, OECD AIGO
Stephanie Ifayemi (moderator) | Head of Policy, Partnership on AI

Main Takeaways

IMPLEMENT SAFETY TESTS AND EVALUATIONS THAT START AT THE DESIGN PHASE:
A consensus emerged around the urgent need for robust safety tests and evaluations for AI systems from an early stage. Speakers stressed the importance of mitigating risks through thoughtful design and regulation and the need for third-party model assessments.

ENGAGE A BROAD STAKEHOLDER COMMUNITY IN THE DEVELOPMENT OF MODEL EVALUATIONS, RISK THRESHOLDS, AND THIRD-PARTY ASSESSMENTS:
Speakers noted the risk of regulatory capture in model evaluations and risk assessments, which should be developed by a broad group of stakeholders, including civil society representatives.

FORMALIZE A COMMON DEFINITION OF HIGH-RISK OR FRONTIER AI MODELS:
Solaiman and Dr. Barnhart highlighted the difficulties in establishing clear thresholds and criteria for AI models. This includes the challenges in distinguishing between different types of models and the subjective nature of model evaluation. There must be an inclusive, interdisciplinary, and ongoing process to develop a definition of high-risk or frontier AI models.

SET ENFORCEABLE MECHANISMS FOR AI SAFETY:
Governments have a responsibility to operationalize AI safety through the investment and implementation of mechanisms that ensure that AI systems are safe. Specifically, governments should establish AI safety institutions and invest in sociotechnical risk-mitigation tools and evaluations.

Discussion

Policymakers across jurisdictions have been asking themselves which tools they should employ to evaluate safety and security in AI systems. This session brought to light the varied and complex aspects of managing the safety and security of foundation models, stressing the need for collaborative and multifaceted approaches involving regulation, policy frameworks, technical solutions, and stakeholder engagement.

Drawing from his technical expertise, Prof. Goldstein reminded the audience that the ability to jailbreak AI systems does not inherently signal a lack of security. Although all systems are vulnerable to breaches, two dimensions of risk prevention must be operationalized across the industry: **precautionary measures applied in the design and development stages, and a contextual approach toward risk assessment at the application level**, to cover risks related to deployment. To increase safety and security measures in the design phase of AI systems, Prof. Goldstein suggested learning from similar risk management strategies applied to other, more traditional, softwares. In doing so, policymakers should bear in mind that foundation models, if compromised, could have extensive negative impact, due to their widespread use in applications across domains.

Prof. Goldstein outlined strategies to mitigate harms associated with AI, beyond technical solutions at the laboratory level. Notably, he called for **platform moderation—the detection and labeling of generative AI content**, to foster public awareness of content authenticity and AI systems' capabilities. He suggested tech companies should be proactive in implementing those measures, with a **consortium of major players in the information ecosystem**.

Dr. Joslyn Barnhart echoed Prof. Goldstein's remark on the unavoidable nature of adversarial attacks in AI. She pointed out the need for society-wide consensus in balancing the benefits and risks of AI technologies. As consensus emerges around the most dangerous risks, we must identify models likely to cross those red lines through a set of specific criteria and apply rigorous scrutiny through **sociotechnical assessments and safety tests for foundation models**.

Analyzing concrete measures to establish red lines, Dr. Barnhart highlighted the challenges licensing may pose to new market entrants and stressed the centrality of model evaluation in AI policy. She underscored **the need for academics and civil society to contribute to inclusive third-party assessments, especially of foundation models**. Finally, Dr. Barnhart acknowledged **the increasing role of governments in AI governance**, motivated by public demand and industry's need for legal clarity. She stressed governments' responsibility to invest in safety and public education.

Solaiman pointed out the current inadequacies in evaluation techniques and the lack of consensus on definitions of "safety" within the AI community ... [she] called for more comprehensive criteria for risk assessment of large models ... relying on computational power as a risk threshold, though useful and important as a first step, will be insufficient in the long run.

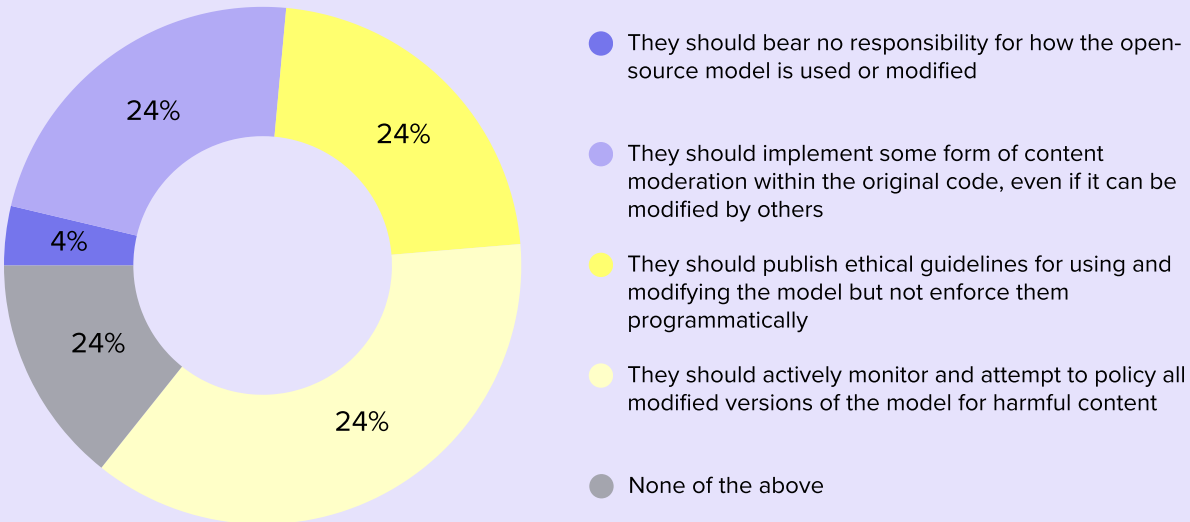
Irene Solaiman focused on the challenges of misuse and unintentional misuse in AI models. She drew a critical distinction between models accessible through APIs and models with open weights, noting the distinct risks each type presents. Solaiman underscored the difficulties in establishing clear risk thresholds for these models and called for extensive work to **build robust policies spanning the gradient of model release methods, from proprietary to open-source**. Meanwhile, we should also implement specific policies to govern the use of generative AI to preserve academic integrity.

She pointed out the current inadequacies in evaluation techniques and the lack of consensus on definitions of "safety" within the AI community. Advocating for a collaborative approach to tackling unknown risks, Solaiman called for more comprehensive criteria for risk assessment of large models, encompassing sociotechnical aspects. Relying on computational power as a risk threshold, though useful and important as a first step, will be insufficient in the long run.

Juraj Čorba provided insights into the evolving AI policy landscape, highlighting the shift from industry-led AI narratives to governmental initiatives in defining AI governance. Čorba expressed concerns over **regulatory capture in certain jurisdictions if big tech companies are allowed to set standards for foundation AI models**. Analyzing the best approach to governing the safety and security risks of foundation models, he drew a parallel with the crypto sector, suggesting that **emerging technologies should be integrated into existing regulatory frameworks, rather than completely revamping them**.

Čorba also discussed the role of voluntary commitments in AI governance. Although valuable in initiating discussions, they are undoubtedly insufficient for holistic and effective governance. He highlighted the varying approaches to AI governance across different jurisdictions and the importance of considering both technological and societal factors. Čorba called for a shared commitment across jurisdictions to adopt a proactive stance to AI governance: rather than focusing solely on technology, **policies should also influence societal behaviors and values to steer AI development towards the common good**.

In 2023, Mistral AI, a French AI startup, released an open-source language model that will provide detailed instructions for suicide, killing one's spouse and acquiring class-A drugs. Which of the following responsibilities should apply to developers?



FIRESIDE CHAT | **DHS Priorities for AI Governance**

WATCH NOW

Robert Silvers | Under Secretary, U.S. Department of Homeland Security
Nicolas Miailhe (moderator) | Founder, The Future Society

Main Takeaways

DEPLOY PERIODIC IMPACT ASSESSMENTS AND OVERSIGHT PROCESSES TOWARD THE USE OF AI BY LAW ENFORCEMENT:
AI technologies can be leveraged to improve the performance of law enforcement in fulfilling their statutory obligations. Considering the level of uncertainty and lack of regulation of AI technologies presently, establishing oversight mechanisms with direct participation of civil society is crucial for fostering trust in public institutions and increasing democratic resilience.

DRIVE TALENT AND INVESTMENT FOR LAW ENFORCEMENT TO FIGHT THE USE OF AI IN CRIMINAL CYBER ACTIVITIES:
Law enforcement agencies require cutting-edge technical tools and expertise to develop efficient strategies to curb the rapidly expanding use of AI in criminal activities. Governments must direct funding and talent to those efforts, while also ensuring strict oversight of agencies' use of AI.

INCLUDE CIVIL SOCIETY REPRESENTATIVES IN GOVERNANCE BOARDS AT LAW ENFORCEMENT AGENCIES:
Participation and decision-making power in governance boards allow for civil society to influence internal operations, practices, and policies related to the use of AI, including the power to establish red lines. This would increase the transparency of law enforcement activities and contribute to social acceptance of AI's use in law enforcement.

Discussion

National security and law enforcement institutions around the world encounter a double-edged sword in responding to AI's impact: while it enables new, large-scale threats to countries and their populations, it also presents state forces with sophisticated technological tools that could facilitate the fulfillment of their mandate. This fireside chat, moderated by The Future Society's founder, Nicolas Miailhe, featured insightful remarks from the Department of Homeland Security's Under Secretary, Robert Silvers, on the evolving role of AI in cybersecurity and governance.

Under Secretary Silvers highlighted the increasing use of AI to automate cyber attacks, with sophisticated techniques that make it more dificult for law enforcement to detect scams. Conversely, AI is also a powerful tool for cyber defense, offering innovative ways to protect against these advanced threats. This dual role underscores an emerging arms race in the cyber domain, where both attackers and defenders leverage AI capabilities.

Addressing the border-agnostic nature of digital security challenges, Under Secretary Silvers stressed the importance of **international collaboration in AI governance and regulatory harmonization**. To ensure consistency in protecting private data and securing reliable networks in global operations, companies must align their operations with diverse regulatory regimes.

A unified global response to AI challenges would alleviate the burden of cross-border operations, benefit companies, and improve security across the value chain.

Under Secretary Silvers also discussed the Biden administration's efforts to **harness AI for public safety, including detecting illegal substances and products made with forced labor through AI-enabled supply chain mapping**. He emphasized DHS's commitment to responsible AI use, ensuring privacy, bias mitigation, and civil rights are central to algorithmic decision-making.

Finally, Under Secretary Silvers emphasized the need to **transform voluntary industry commitments into codified regulations, policies, or treaties**. He highlighted the formation of DHS's Artificial Intelligence Safety and Security Board, a blend of federal leads, industry experts, and academics tasked with developing best practices for AI safety and security.

Under Secretary Silvers highlighted the increasing use of AI to automate cyber attacks ... Conversely, AI is also a powerful tool for cyber defense, offering innovative ways to protect against these advanced threats ... This dual role underscores an emerging arms race in the cyber domain, where both attackers and defenders leverage AI capabilities.

ROUNDTABLE DIALOGUE | Navigating AI Deployment Responsibly: Open-Source, Fully-closed, and the Gradient in Between

Alyssa Ayres | Dean, George Washington University Elliott School of International Affairs
Nicolas Miailhe | President and Founder, The Future Society
Luis Aranda | AI Policy Analyst, OECD.AI
Anthony Aguirre | Executive Director & Secretary of the Board, Future of Life Institute; Professor of Physics, University of California, Santa Cruz
Russell Wald | Deputy Director, Stanford Institute for Human-Centered AI (HAI)
Elizabeth Seger | Research Scholar, Centre for the Governance of AI (GovAI)
Peter Cihon | Senior Policy Manager, GitHub Heather Frase | Senior Fellow, Georgetown's Center for Security and Emerging Technology (CSET)
Ian C. Haydon | Science Communicator, Institute for Protein Design

The practice of "open-sourcing" technologies has been a subject of both admiration and criticism. On the one hand, it has allowed for "democratization" and inclusivity in technological developments, and for software robustness through community-driven inspection and audits, red-teaming, and bug detection. On the other hand, it allows for these technologies—harboring unknown and potentially hazardous capabilities—to be more readily misused. As AI systems become more capable, the potential for their misuse and harm, such as risks to cybersecurity and biosecurity, grows correspondingly.

This interactive roundtable dialogue brought together over 100 AI policy experts to brainstorm actionable recommendations for adapting release strategies for powerful AI systems.

The speakers presented short remarks contextualizing the state of AI research and practices and the role of open-source in the AI ecosystem. These remarks were then followed by group discussions and a debriefing session.

Several speakers challenged the notion of a binary between "open" and "closed" models, pointing toward a spectrum of options regarding the level of access to system components such as datasets, code, model cards, and model weights. Given their widespread use and potential for both benefit and harm, the release strategies of recently developed large language models were compared. Biological design tools, which offer groundbreaking medical solutions but also present biosecurity risks, were also discussed as a use case of interest.

> *On the one hand, ["open-sourcing"] has allowed for "democratization" and inclusivity in technological developments, and for software robustness through community-driven inspection and audits, red-teaming, and bug detection. On the other hand, it allows for these technologies ... to be more readily misused.*

Discussions probed into the jurisdictional challenges of governing the release of models. Participants acknowledged that **wide sharing of model weights can make it difficult, if not impossible, to trace and attribute instances of misuse, and thereby seek redress in such cases**. Some pointed out that transparency does not necessarily have to mean granting full access to the model, but stressed that closed models must also be expected to adhere to rigorous transparency requirements, including assessments by third parties. **Some discussants saw promise in a risk-based approach, combining national mechanisms such as licenses and global UN-sanctioned certification, to regulate the deployment of closed models with potential for tangible harmful outcomes**.

Discussions underscored the importance of considering a liability framework based on the capabilities and generality of AI systems. **Licensing emerged as a key mechanism, with some discussants proposing a centralized authority or a consortium for overseeing a model testing process prior to open-source release. Some discussants also stressed that the global majority should be appropriately represented in such governance processes**. The idea of an international mechanism, possibly akin to a CERN for AI, was proposed, focusing on beneficial applications and establishing a new social contract with internationally accountable governance.

Suggested elements toward more robust governance of open-source AI included external expert-led red-teaming, government-funded audits, and incident reporting.

REMARKS | **Audrey Plonk**

WATCH NOW

Audrey Plonk | Head of Division, Digital Economy Policy, OECD

Audrey Plonk provided remarks focused on recent developments in AI safety and the OECD's dedication to international coordination in AI governance. In the past few months, the organization took part in key forums, such as the G7 ministerial meeting on the Hiroshima AI process, the UK AI Safety Summit, and its own multistakeholder network of AI experts.

Plonk observed that **AI safety has transitioned from a specialized technical concern to a top priority for governments worldwide**. This shift has sparked debates on the necessity of an international governance regime for advanced foundation models. In this sense, comparisons with institutions like CERN, the IAEA, and the IPCC have been increasingly drawn.

Regardless of the form an international institution may take, **international norms remain crucial to promote AI safety, robustness, trustworthiness, and human rights**. While the OECD AI Principles lay a foundational framework, she acknowledged the need for additional measures as AI technologies proliferate. Plonk stressed that the OECD is developing responsible business conduct guidelines for AI, aiming for flexible yet enforceable mechanisms to guide AI companies operating internationally and address AI-related disputes through mediation.

Additionally, Plonk highlighted the launch of the OECD AI Incidents Monitor, a tool to monitor global news in real time to detect and classify AI-related incidents, offering a vital resource for international risk management and data-driven policymaking

MEASUREMENT & STANDARDS

KEYNOTE | **Dr. Erwin Gianchandani**

WATCH NOW

Dr. Erwin Gianchandani | Assistant Director for Technology, Innovation and Partnerships, U.S. National Science Foundation

Dr. Gianchandani presented the National Science Foundation's (NSF) role in driving AI innovation in the US. He highlighted how the NSF is advancing its mission with the new directorate for technology, innovation, and partnerships, aimed at equipping researchers, startups, and entrepreneurs with resources to translate ideas into societal benefits.

Dr. Gianchandani noted that accelerating research is key to leveraging AI's transformative potential responsibly. **AI models' escalating capabilities have the potential to accelerate scientific discoveries, provide solutions to societal challenges, and reshape how we interact with technology**. Dr. Gianchandani stressed NSF's leadership in the pilot implementation of NAIRR (National AI Research Resource) to expedite resource accessibility for the research community to address those societal challenges. In addition, he outlined the NSF's efforts in funding foundational AI research and its dedication to addressing current and future risks.

Collaborative and interdisciplinary partnerships are at the heart of NSF's approach to AI governance. The Foundation has collaborated with NIST in establishing the Institute for Trustworthy AI in Law and Society (TRAILS)—a co-host of The Athens Roundtable—and created the National AI Research Institutes program

PANEL | Decoding AI: Challenges in Classification, Measurement, and Evaluation

WATCH NOW

Elham Tabassi | Associate Director, Information Technology Laboratory and Chief AI Advisor, U.S. NIST
Jared Mueller | Head of External Affairs, Anthropic
Sebastian Hallensleben | Chair of JTC 21, CEN, CENELEC
Emmanuel Kahembwe | CEO, VDE UK
David Broniatowski (moderator) | Associate Professor, The George Washington University; Co-PI and GW Site Lead, NIST-NSF TRAILS

Main Takeaways

BROADEN THE SCOPE OF AI EVALUATIONS TO INCLUDE SOCIETAL ROBUSTNESS AS A KEY METRIC:
Governments must foster interdisciplinary approaches focused on the safety and societal implications of AI systems. Ensuring that AI systems are developed and deployed with a comprehensive understanding of their wider impacts will only be possible with a broader pool of stakeholders and impacted communities participating in standard-setting. It's crucial that this work be developed in coordination with various AI safety institutes globally to share and implement best practices.

DEVELOP UNIFORM METRICS, METHODOLOGIES, TRANSPARENCY REQUIREMENTS, AND REPORTING STANDARDS TO FACILITATE THE COMPARISON AND ASSESSMENT OF AI SYSTEMS ACROSS DIFFERENT DOMAINS:
Speakers highlighted how crucial interoperability is in standardizing evaluation processes and making them more transparent and effective.

ALLOCATE RESOURCES TO STANDARD-SETTING COMMITTEES TO ENSURE BROADER PARTICIPATION FROM A DIVERSE ARRAY OF STAKEHOLDERS:
This approach would enable more equitable representation and input in the standard-setting process, including with academia and civil society representatives, ensuring that the standards developed are reflective of a wider range of perspectives and needs. Inclusion is crucial given the cross-jurisdictional deployment of AI models and their disproportionate impact on the global majority.

Discussion

Definitions, metrics, benchmarks, and evaluations play a crucial role in the governance of advanced AI systems. In this session, AI experts delved into established and emergent challenges in classification, measurement, and evaluation, proposing concrete measures to achieve scientifically credible and robust tools and processes for AI governance.

Sebastian Hallensleben opened the discussion by exploring the evolving nature of AI terminology, highlighting the lack of agreed-upon definitions for terms like "foundation models" and "generative AI." He emphasized **the importance of differentiating between raw models like GPT-4 and more application-oriented systems like ChatGPT, noting how these distinctions influence AI governance.** He called for a common understanding of concepts like trust, truth, and facts, especially in the context of generative AI's impact on consumer applications and societal challenges.

Elham Tabassi emphasized the evolution of NIST's approach to AI measurement and evaluation, particularly following the comprehensive Executive Order 14110 from October 2023. She highlighted **NIST's role in developing guidelines for evaluating potentially harmful AI systems, including red-teaming strategies, and in creating test environments in collaboration with other agencies.** Tabassi pointed out that current AI system evaluations primarily focus on technical robustness—a relatively urgent priority. She stressed that **methods should assess AI systems in their real-world contexts in a scientifically accurate and reproducible manner,** acknowledging the complexity of today's technology.

Emmanuel Kahembwe added to this discussion by emphasizing the limitations of the current training of technical AI experts. Such professionals often receive training focused on a narrow set of systems (often limited to those that they develop and deploy), with an emphasis on technical performance metrics. He further noted that **governments should facilitate coordination between AI Safety Institutes to share and implement best practices.**

Jared Mueller addressed the scrutiny required to ensure the safety of large AI models, acknowledging that while computational resource utilization (floating-point operations per second, or “FLOPS”) may not be a perfect measure, it currently serves as the best available standard. He also highlighted **the risk of regulatory capture and the need for diverse expertise in evaluating large models**. Mueller underscored the importance of including a broad range of specialists—from civil society to government experts—beyond governance and policy professionals, to comprehensively cover the expanding risk profiles in the field of AI.

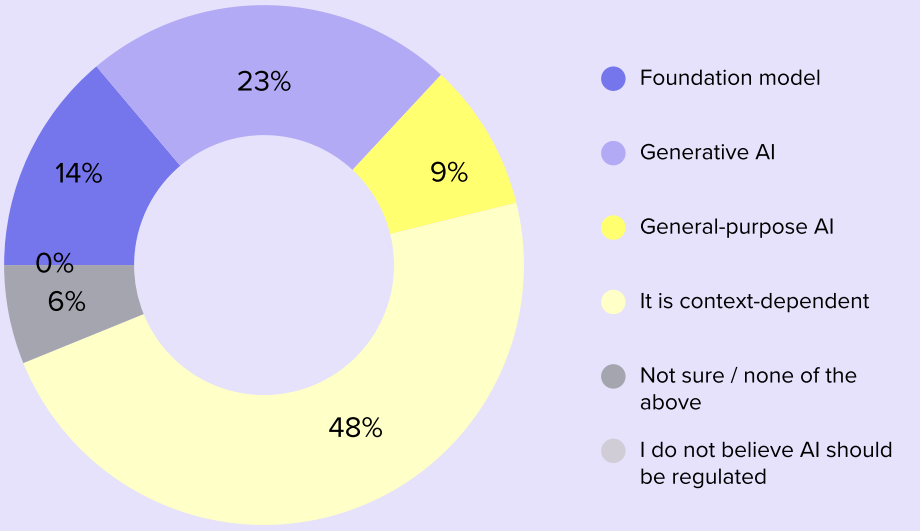
Mueller underscored the importance of including a broad range of specialists—from civil society to government experts—beyond governance and policy professionals, to comprehensively cover the expanding risk profiles in the field of AI.

Delving into the intricacies of consensus-building among diverse stakeholders, **speakers highlighted the need to broaden the range of expertise and backgrounds involved in standard-setting**, recommending the inclusion of communities impacted by AI technologies. Integrating diverse insights from the onset would contribute to more holistic and impactful AI standards. Drawing from his role at CEN-CENELEC Joint Technical Committee 21 on AI (JTC21), Dr. Hallensleben stressed that these committees bear the responsibility of actively reaching out to ensure diverse participation. While it is challenging to achieve consensus with a large and diverse pool of stakeholders, **a diversity of perspectives tends to enhance the quality and applicability of the standards.** In this sense, standards committees based in the Global North should not overlook the need for representation from the Global South if they aim to have international applicability.

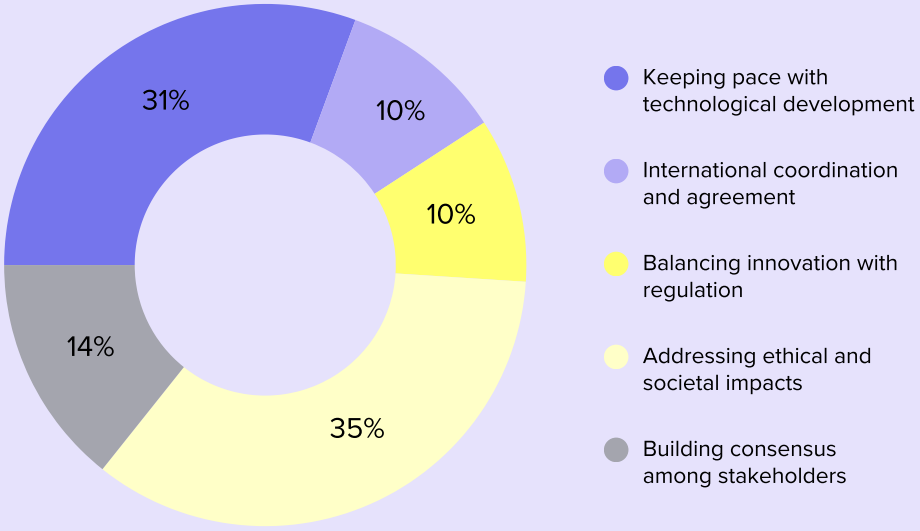
Looking at practical challenges, speakers identified the voluntary nature of participation as a critical barrier to inclusivity in standards-setting. Stakeholders, particularly those most impacted by AI developments, often lack the resources to engage in voluntary standard-setting processes. To address this gap, Kahembwe proposed **reevaluating the voluntary aspect of standards-setting activities and allocating resources to remunerate participants**. Furthermore, speakers emphasized the importance of effectively **translating consensus into clear, technically useful documentation**. This approach ensures that AI ethics standards are not only comprehensive and representative but also practically useful for programmers and engineers.

Finally, **speakers analyzed the role of standards in shaping not only industry practices but also legal outcomes in cases involving AI technologies.** As standards gain strength and legitimacy, they could increasingly play a pivotal role in judicial cases and arbitration. Legal practitioners and judges might be more inclined to rely on these standards in their rulings and to consider expert witnesses familiar with these benchmarks. This potential judicial reliance on standards underscores the need for them to be well-established, legitimate, and reflective of broad expert consensus.

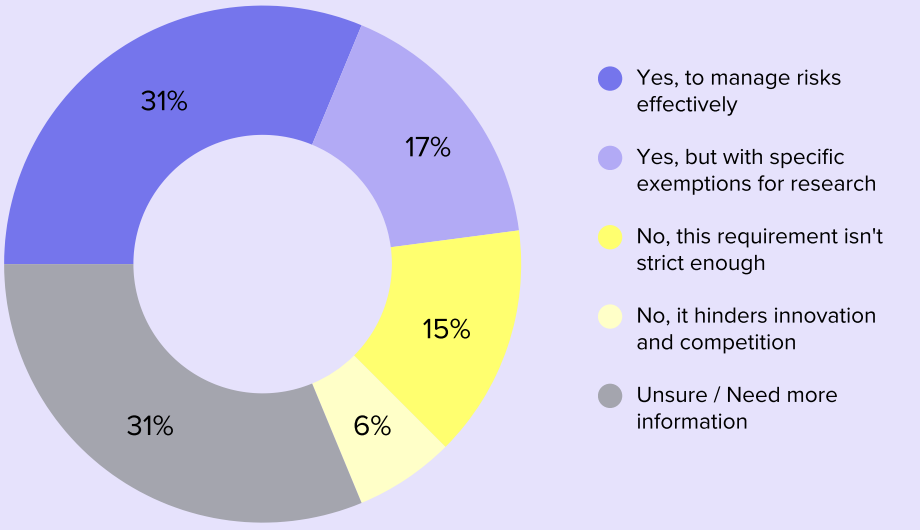
Which term do you believe is most appropriate as the 'object of regulation' for laws and guidelines concerning advanced AI systems?



What do you expect to be the biggest challenge in developing standards for AI systems?



Do you approve of the Biden Administration's Executive Order mandating that developers of foundation models that exceed a compute threshold must submit detailed reports to the US government?



REMARKS | **John C. Havens**

WATCH NOW

John C. Havens | Regenerative Sustainability Practice Lead, The IEEE Standards Association

John C. Havens provided remarks on **leveraging AI for long-term human and planetary well-being.** He noted that inclusion, sustainability, and equal opportunity are at the core of long-term human flourishing. This perspective is notably reflected in the UN Sustainable Development Goals (SDGs) and the OECD Better Life Index. Drawing from those initiatives, Havens underscored that, in the age of AI, it's crucial to extend our developmental understanding beyond traditional economic metrics such as GDP.

Havens highlighted IEEE's role in steering AI governance in that direction, referencing the IEEE 7010 standard for well-being impact assessment of AI systems (2020) and the pioneering work on Recommended Practice for the Provenance of Indigenous Peoples' Data. Furthermore, IEEE developed the Planet Positive 2030 program reflecting a commitment to regenerative sustainability, which seeks to foster a net positive impact on the planet.

Highlighting the urgent need to protect younger generations and their future, Havens advocated for **new standards in age-appropriate design and sustainability, emphasizing the inclusion of children and future generations in technology innovation.** Havens concluded by challenging the predominance of Western rationality in AI governance, and advocating for values like relationality and community care to guide AI development

REMARKS | **Margot Skarpeteig**

WATCH NOW

Margot Skarpeteig | Program Manager, Human Rights, Empowerment and Inclusion, The World Bank

Margot Skarpeteig reflected on the 75th anniversary of the Universal Declaration of Human Rights against the backdrop of significant technological advancements, particularly in AI. She emphasized the challenges posed by these advancements, noting how **the potential of AI as a force for positive change is currently overshadowed by threats to human dignity and agency.**

Skarpeteig underscored The World Bank's awareness of its crucial role in upholding human rights within the global digital marketplace.

She highlighted the efforts of the Human Rights Trust Fund, which supports World Bank staff in understanding the intersection of human rights and development in their operations and analytics.

Furthermore, Skarpeteig discussed The World Bank's initiative to develop a comprehensive framework for identifying and mitigating the human rights risks associated with AI in their institution's operations.

REGULATION & ENFORCEMENT

KEYNOTE | U.S. Senator Richard Blumenthal

WATCH NOW

Richard Blumenthal | United States Senator

Reflecting on the burgeoning influence of AI in 2023, Senator Blumenthal underscored the significant impact AI has on the economy, safety, and democracy. **He cautioned against Congress repeating past errors seen in the technological revolutions of the previous decade,** particularly referencing the challenges faced with the rapid growth of social media. Senator Blumenthal noted Congress's failure to act in the past, which led to the rise of monopolistic companies wielding disproportionate power.

Drawing from his experience as chair of the Judiciary Subcommittee on Privacy, Technology, and the Law in 2023, he shared insights from witness testimony by industry leaders—including OpenAI CEO Sam Altman, Anthropic CEO Dario Amodei, and Microsoft President and Vice Chair Brad Smith—who, **unlike social media executives in the past, expressed a unanimous call for AI regulation**.

In August, Senator Blumenthal and Senator Hawley announced a **bipartisan framework for a U.S. AI Act**. This framework proposes establishing a **licensing regime for entities engaged in high-risk AI development and creating an independent oversight body** with AI expertise. It lays out specific principles for upcoming legislation aimed at protecting national and economic security, enforcing transparency about AI model limitations and uses, protecting consumers and children, and implementing rules like watermarking, disclosure of AI usage, and data access for researchers. Furthermore, **the framework addresses the accountability of AI companies, holding them liable for privacy breaches, civil rights violations, or other harms**.

Reflecting on international developments, Senator Blumenthal highlighted the significance of the EU's AI Act, lauding it as a groundbreaking effort that sets baseline rules and standards for AI, and providing a valuable model for AI regulation akin to the EU's initiatives in privacy, competition, and online safety.

KEYNOTE | **U.S. Senator Brian Schatz**

WATCH NOW

Brian Schatz | United States Senator

Senator Brian Schatz's keynote addressed the critical issue of **regulating dual-use foundation models at the federal level in the United States.** He emphasized the vital role of the federal government in this endeavor while acknowledging the current lack of a unified approach.

Senator Schatz critiqued traditional regulatory methods, which typically either address harms on a case-by-case basis or establish an extensive list of statutory provisions, which would be ineffective for the rapidly evolving field of AI. He stressed **the need for the US to develop basic, common-sense, future-proof principles that encourage developers and deployers to innovate responsibly.**

Stressing the crucial role of enforcement, Senator Schatz highlighted the role of federal agencies, but cautioned against oversimplified statutory frameworks that could be manipulated by tech corporations. Senator Schatz stressed the importance of a nuanced and adaptable regulatory framework capable of addressing the multifaceted challenges posed by AI technologies.

Focused on the immediate steps necessary in AI regulation, Senator Schatz proposed requiring clear disclosure when online content is machine-generated, which would enhance transparency and accountability in the digital realm. Furthermore, he emphasized the urgent need for **regulations concerning the use of data in training AI models,** advocating for a duty of care from data collectors towards individuals whose data is being utilized.

KEYNOTE | U.S. Senator Amy Klobuchar

WATCH NOW

Amy Klobuchar | United States Senator

Legislative guardrails are essential not only to safeguard consumers and intellectual property but also to preserve the very foundations of democracy. Senator Amy Klobuchar's keynote underscored the urgent need for legislative guardrails for generative AI and the importance of international coordination.

Increasingly sophisticated AI-generated content can spread misinformation related to elections, such as inaccurate information about voting logistics, posing concrete risks to democratic processes and the upcoming election in the United States. Senator Klobuchar highlighted key bipartisan efforts to **combat the growing threat of deepfakes in U.S. electoral processes**. She discussed the need to confront deceptive practices while ensuring free speech—an approach encapsulated in the **Deceptive AI Act**, aimed at **curbing the use of fraudulent content in political advertising**.

Highlighting another critical issue with regulating generative AI and protecting the information ecosystem, Senator Klobuchar advocated for the **protection of individuals and content creators** against the unauthorized use of their voice, likeness, and proprietary work. She stressed the importance of protecting local news organizations, for instance, from undue reproduction and use of training data without compensation by large platforms. The **Journalism Competition and Preservation Act**, as she mentioned, aims to empower local news outlets to negotiate fair compensation for their content—an issue closely related to information integrity and trust in information ecosystems and democratic institutions.

Taking the discussion back to power dynamics and democratic control over AI, Senator Klobuchar emphasized the need to **modernize U.S. competition laws to address the unique challenges posed by the concentration of power in the AI landscape and called for legislation to ensure transparency and accountability, particularly for high-risk AI applications**. Finally, recognizing that AI's challenges transcend national borders, Senator Klobuchar advocated for **global cooperation in developing and harmonizing AI governance frameworks** to effectively address these universal challenges.

FIRESIDE CHAT | **Regulating AI across its value chain**

WATCH NOW

Addie Cooke | Global AI Policy Lead at Google Cloud, Google
Cameron Kerry | Ann R. and Andrew H. Tisch Distinguished Visiting Fellow, Center for Technology Innovation, Brookings Institution; Former General Counsel and Acting Secretary, U.S. Department of Commerce
Anna Gressel (moderator) | Counsel, Paul, Weiss

Main Takeaways

BALANCE HORIZONTAL FRAMEWORKS WITH SECTOR-SPECIFIC REGULATION:
Stakeholders must advance discussions around the legal considerations in allocating liability along the AI value chain to develop robust and legally sound doctrine and policy. Emerging AI liability regimes should consider existing regulatory frameworks and, when appropriate, complement them, such as with contractual, legal, and regulatory liability in different sectors.

DEVELOP INTERNATIONAL STANDARDS AND MECHANISMS FOR INCREASED CORPORATE TRANSPARENCY:
Speakers converged on the need for industry-wide standards independent of binding regulation, alongside investments in model monitoring tools, transparency requirements, incident reporting protocols, and auditing by independent third parties, to ensure AI development and deployment and is ethical and preserves public trust.

Discussion

Liability, often defined in contracts between value chain actors, is increasingly being considered at the regulatory level as the consequences of AI systems grow more severe. This fireside chat delved into the complexities of regulating AI across its value chain.

Acknowledging that regulation will be crucial for both risk mitigation and industry innovation, this discussion highlighted the significant challenge of allocating responsibility within the AI value chain. Discussions spanned governance approaches for accountability and liability, how supply chain actors are reacting to the regulatory trends, and the balancing act of advancing responsible AI across jurisdictions.

Speakers emphasized **the urgency of aligning risk assessment and compliance with regulatory trends, as the cost of non-compliance increases** with every major jurisdiction that enacts AI regulations. A key point of discussion was the regulation of AI developers and deployers and the varying levels of risks associated with different sizes of AI models, particularly in the context of generative AI applications. The conversation touched upon how the EU AI Act might influence regulatory approaches to foundation models in other jurisdictions.

Moderator Anna Gressel touched upon the evolving nature of liability in the AI sector. Beyond regulation of dual-use foundation models, **product liability should also be considered in the US, following Europe's lead on the matter**. Given the current scenario of divergent regulatory proposals and self-governance approaches, panelists underscored the importance of developing and implementing standards, as set by organizations like ISO and IEEE, to harmonize approaches and reduce compliance costs across jurisdictions.

Focusing on corporate responsibility, Cameron Kerry **highlighted the need for companies to invest in transparency, incident reporting, and thorough auditing processes regardless of binding regulation.** Kerry drew an analogy to the need for careful planning and diligent, repeated measurement in the practice of carpentry, emphasizing the need for diligence and precision in AI regulation and deployment.

Addie Cooke pointed out the increasingly relevant role of **model monitoring tools in the industry, which can help raise the technical bar for risk assessment**. She noted that evaluations should be done at different stages of the value chain and praised NIST's Risk Management Framework for its adaptability and usefulness for the industry.

Given the current scenario of divergent regulatory proposals and self-governance approaches, panelists underscored the importance of developing and implementing standards, as set by organizations like ISO and IEEE, to harmonize approaches and reduce compliance costs across jurisdictions.

FIRESIDE CHAT | **Coordinated approaches for AI governance**

WATCH NOW

Dragos Tudorache | Member of the European Parliament (pre-recorded remarks)
Lynne E. Parker | Associate Vice Chancellor and Director of the AI Tennessee Initiative, University of Tennessee, Knoxville
Marek Havrda | Deputy Minister for European Affairs, Office of the Government of the Czech Republic
Nicolas Moës (moderator) | Director, European AI Governance, The Future Society

Main Takeaways

DEVELOP AND IMPLEMENT A SET OF REGULATORY TOOLS TO OPERATIONALIZE SAFETY BY DESIGN: Such tools must be interoperable across jurisdictions, given the borderless character of the foundation models value chain. Regulators should invest in regulatory sandboxes to rigorously test and refine foundation models pre-deployment. This effort should be informed by comprehensive regulatory guidance, global metrics, industry-wide standards, and interoperable benchmarks.

STRENGTHEN CROSS-BORDER INFORMATION-SHARING BETWEEN REGULATORS: This is key to harmonize approaches to AI governance between the EU, US, and other global partners. This effort should include sharing best practices, and knowledge critical for tackling challenges related to enforcement.

ENHANCE INVOLVEMENT OF DIVERSE STAKEHOLDERS IN AI GOVERNANCE TO GATHER ROBUST EVIDENCE ABOUT AI'S IMPACT: Inclusion can be operationalized through advisory panels, public forums led by civil society, and other methods of obtaining continuous feedback from underrepresented communities.

Discussion

Over the course of 2023, increasing market demand for AI has pushed the public interest to the margins. However, regulatory developments have provided a democratic path to balance stakeholders' power and protect the public interest: the European Union's AI Act. In its final stages, it represents a robust effort to regulate AI models increasingly prevalent in consumer markets and impose guardrails that uphold safety and fundamental rights. Nevertheless, ongoing work is necessary to ensure the Act's strength and enforceability and to transmit regulatory lessons to other jurisdictions. The role of institutions responsible for enforcement at the national level—such as the EU AI Act's European AI Office—is crucial in this regard. This fireside chat focused on **the critical role of coordination between AI policy enforcement bodies,** particularly across those of the EU and the US.

In opening remarks to the panel, MEP Dragos Tudorache shared insights on **the trilogue process for the EU AI Act**, underscoring the importance of learning from the EU regulatory journey and reflecting on upcoming challenges for enforcement. **The international community must ramp up a coordinated approach to maximize countries' capacity for enforcement,** be it within the EU AI Act jurisdictions or beyond EU borders as regulations emerge in other countries.

Dr. Lynne Parker offered insights into **how the EU's regulatory path might have influenced the U.S. government's approach to AI governance**. When it comes to narrow systems, the sectoral-based approach discussed in the EU resonates with the US regulatory structure, comprising different agencies with expertise in different economic sectors. Those agencies are already studying or investigating the impact of AI within their mandates. Dr. Parker suggested that **federal institutions are well-equipped to take up a two-pronged approach: sector-specific regulations coupled with a comprehensive AI governance framework**, such as the work the U.S. executive branch has been advancing since the publication of the **Blueprint for an AI Bill of Rights** and, more recently, **Executive Order 14110**.

Moving the discussion from executive powers to legislative powers, Marek Havrda articulated the challenges in transitioning AI legislation from theory to practice, with a particular focus on the role of a **European AI Office**. This institution would have a central role in gathering intelligence around the regulatory learning stemming from the oversight in member states' jurisdictions and from regulatory sandboxes—should the provisions be approved in the final text of the EU AI Act. Finally, looking into the Executive's role, Havrda underscored **the importance of coordination among national AI offices**

> *Discussions tend to focus on the downstream impact of AI applications on society, such as with the deployment of surveillance technologies, but, speakers remarked, it is urgent to rein in corporations' actions during the design and development of AI systems, rather than focusing solely on deployment.*

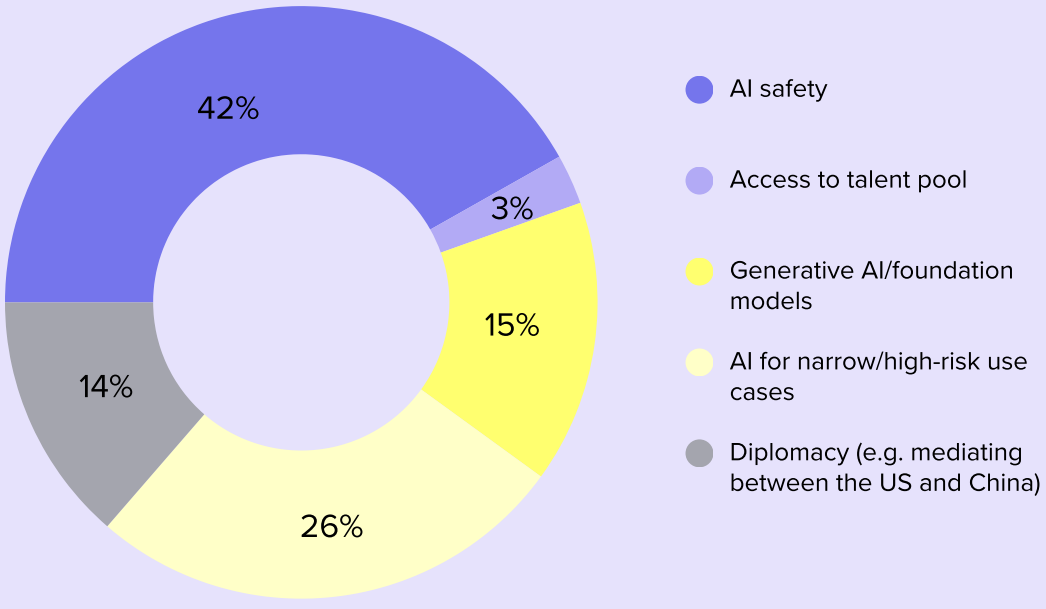
for consistent enforcement across the EU, especially with respect to high-risk systems. If successful, he remarked, this model could be extended to international collaborations.

Shifting the lens to the global stage, the **G7 Hiroshima Code of Conduct** was identified as a significant step in guiding the behavior of foundation model developers. However, there remains an **urgent need for binding regulations**, like the EU AI Act, to manage and mitigate systemic risks effectively.

As we broaden the debate around AI governance and enforcement to **include underrepresented voices and increase democratic participation**, speakers discussed that we must be cautious about overshadowing complex yet crucial AI discussions. Foundation models have profound implications on a wide array of downstream applications affecting the global population. Discussions tend to focus on the downstream impact of AI applications on society, such as with the deployment of surveillance technologies, but, speakers remarked, it is urgent to **rein in corporations' actions during the design and development of AI systems**, rather than focusing solely on deployment.

As the world turns its eyes to the profound and borderless impact of some foundation models, the concept of **"safety by design"** is crucial in mitigating systemic risks at the initial stages of AI development..

In which area is europe most likely to shape the future of AI?



Remarks by Co-hosts

REMARKS | **Ambassador Ekaterini Nassika**

WATCH NOW

Ekaterini Nassika | Ambassador of the Hellenic Republic to the USA

The Ambassador for Greece in the United States highlighted Greece's leadership in advancing the rule of law in the age of AI. As AI will bring about a transformative moment for humanity, **it will be crucial to leverage its potential to promote stronger democracies**. The Ambassador also acknowledged the challenges accompanying such technology, noting that AI must also be tamed so as not to endanger the values of democracy and the rule of law.

REMARKS | **Stefanos Vitoratos**

WATCH NOW

Stefanos Vitoratos | Co-Founder, Homo Digitalis

Drawing from the Hellenic democratic tradition, Stefanos Vitoratos stressed that the preservation of our fundamental values should be prioritized and reflected in the core of AI development endeavors. Acknowledging the importance of democratic debate, Vitoratos commended the kaleidoscope of perspectives presented by legislators, policymakers, law practitioners, civil society representatives, and developers in the quest to govern AI across jurisdictions. He expressed concern with **the rise of national security discourses that may exclude AI development from public scrutiny**. Vitoratos stressed that stakeholders across jurisdictions hold a common task of forging new mechanisms and institutional solutions to safeguard the rule of law and steer AI development and deployment toward the public interest.

REMARKS | **Dr. Ellen M. Granberg**

WATCH NOW

Ellen M. Granberg | President, George Washington University

The President of the George Washington (GW) University, Dr. Granberg stressed the importance of cross-stakeholder solutions for AI governance, and the critical role conversations such as The Athens Roundtable have in informing both policy and academic priorities in this field. As powerful agents of change, **the role of academics in such conversations is to break down disciplinary silos** to protect and enhance human experience and fundamental rights in the age of AI.

REMARKS | **Dr. Pamela Norris**

WATCH NOW

Dr. Pamela Norris | Vice Provost for Research, George Washington University

Vice Provost Pamela Norris emphasized academia's pivotal role in establishing guardrails and good governance for AI systems, particularly for future generations. Dr. Norris stressed academics' unique role in advising policymakers. She called for rigorous, evidence-based research to inform policy and foster trustworthy AI alongside democratic values. Dr. Norris also underscored **the need for training the next generation of AI professionals to develop AI that is safe and trustworthy**, with the potential to positively transform communities.

> *... Stefanos Vitoratos stressed that the preservation of our fundamental values should be prioritized and reflected in the core of AI development endeavors.*

Conclusion

The fifth edition of The Athens Roundtable shed light on the urgency of adopting a multifaceted approach to AI governance—one that encompasses comprehensive regulations, precise definitions and metrics, and robust enforcement mechanisms.

The recommendations emerging from the dialogue point towards a future where AI development is not only governed by the principles of safety and responsibility, but also steered by a harmonized legal framework that transcends borders and sectors. To achieve this, a collaborative effort is required, bringing together policymakers, developers, civil society, and impacted communities. Beyond coordination, we must develop liability frameworks and governance regimes for general-purpose foundation models that are adaptive and agile. These steps are critical in fortifying our democratic institutions to be resilient to the disruptive potential of AI—ensuring that technological progress does not come at the cost of societal well-being and democratic values.

Moving forward, The Future Society's role in facilitating dialogues and spearheading collaborations for institutional innovation becomes more crucial than ever. The insights and policy recommendations from the Athens Roundtable provide a roadmap for action, but they also serve as a reminder of the challenges ahead. **The goal is clear: guide AI development in a manner that upholds fundamental rights and the rule of law.** Achieving this will require continued commitment, creativity, and cooperation from all stakeholders involved. We look forward to collaborating with Roundtable partners, participants, and readers of this report in furthering our mission of *aligning artificial intelligence through better governance* in the years ahead.

THE
FUTURE
SOCIETY

Contact Us!

GENERAL | info@thefuturesociety.org
PRESS | press@thefuturesociety.org

THE FUTURE SOCIETY
867 Boylston Street, 5th Floor,
Boston MA 02116,
United States

www.thefuturesociety.org