

Policy Backgrounder: Global AI Agreements and Countries' Own Efforts Expose Divergence

May 3, 2024

Trusted Insights for What's Ahead™

A growing number of international agreements seek to align countries' artificial intelligence (AI) safety efforts, yet these efforts expose divergences in approach, potentially challenging further international agreement.

- A partnership announced in April between the new AI Safety Institutes in the US and the UK is the latest in a series of international agreements, which also includes a recent UN AI Resolution and the G-7 AI Code of Conduct.
- The international fora aim to guide each country's efforts in developing national AI policies by seeking alignment on core principles; however, few of these agreements are enforceable.
- The EU's AI Act, which will soon enter into force; countries are forging ahead on establishing AI frameworks, which show wide divergence in approach countries are taking to the growing sector.

Momentum to Align Global AI Efforts

International momentum towards aligning AI efforts continues to grow. On April 1, the new [US AI Safety Institute](#) and its UK peer [AI Safety Institute](#) announced a [Memorandum of Understanding](#) (MOU) to deepen collaboration on AI safety. Signed by US Commerce Secretary Gina Raimondo and UK Technology Secretary Michelle Donelan, the [agreement](#) seeks to align the US and UK's efforts towards building approaches to evaluating advanced AI models, systems, and agents. "Because of our collaboration, our Institutes will gain a better understanding of AI systems, develop, and conduct more robust evaluations, and provide more rigorous and useful guidance," Raimondo said in a [statement](#).

The MOU followed through on November's [AI Safety Summit](#) hosted in Bletchley Park, UK, which brought together the US, China, the EU, and 26 other countries to produce [The Bletchley Declaration on AI Safety](#). While the agreement is non-binding, the participants found agreement on (1) identifying and building shared scientific understanding of AI safety risks, and (2) establishing risk-based policies across parties to the Declaration focusing on transparency for companies developing frontier AI models, creating evaluation metrics and tools for AI systems, and developing public sector capabilities.

"We need global solutions ... and both governments and the private sector need to be part of the solution," [commented](#) Raimondo. "We anticipate that in the weeks and months ahead, more partnerships will help create a global network of AI safety, built through numerous linkages between government-backed scientific institutions." Accordingly, international fora continue making headway.

In March, the UN General Assembly unanimously [adopted](#) the first [global resolution on AI](#), which encourages countries to safeguard human rights, strengthen privacy policies, and monitor AI risks. The non-binding resolution, proposed by the US and co-sponsored by China in addition to 190 other nations, largely focused on key principles rather than specific initiatives. US National Security Advisor Jake

Sullivan [said](#) that while it took nearly four months to negotiate the UN resolution, it gave the world "a baseline set of principles to guide next steps in AI's development and use."

Similarly, in late 2023, the G-7 [agreed](#) on broad AI principles through the Hiroshima Process set up in May 2023 as an AI forum for government leaders. The agreement led to the release of [International Guiding Principles for Advanced AI Systems](#) and an [International Code of Conduct for Organizations Developing Advanced AI Systems](#), outlining governance principles and voluntary guidance for AI developers, which officials [say](#) should act as a bridge until country-specific regulation is put into place.

Global cybersecurity leaders also came to an [agreement](#) last November, as the US, UK, and 16 other countries jointly published [Guidelines for Secure AI System Development](#). The guidelines seek to ensure AI products are "secure by design," recommending responsible AI development and that that models be publicly deployed only after appropriate security testing. The agreement highlighted novel security vulnerabilities of AI systems that rogue attackers could exploit, with the design principles for AI developers acting as a first line of defense to mitigate those risks. Jen Easterly, director of the US Cybersecurity and Infrastructure Security Agency, [stated](#) the guidelines represent "an agreement that the most important thing that needs to be done at the design phase is security."

While global agreements on AI continue to be negotiated, few have direct implications for developers; many still draw from the [2019 OECD AI Principles](#), whose first principle is to achieve "inclusive growth, sustainable development, and well-being" through AI. Instead, they attempt to promote a similar framework as countries work to adopt their own AI laws and standards. International agreements have also sidelined many of the most difficult issues, including those related to copyright, market structure, and environmental impact. The next international forum comes on May 21-22 at the [AI Seoul Summit](#), co-hosted by the UK and the Republic of Korea. Some countries are [expected](#) raise several of these tougher issues in Seoul, yet, many [anticipate](#) significant challenges to achieving further agreement beyond the current principles, highlighting the radically different approaches countries are taking independently.

Countries' Evolving AI Frameworks Show Divergence

United States

The US has taken a decentralized approach, with the Administration [calling](#) on Congress to pass comprehensive AI legislation. As an initial step, the [Executive Order on Safe, Secure, and Trustworthy Development and Use of AI](#) was published in October. The Executive Order provided a comprehensive set of directives to federal agencies to begin developing guidelines for safe AI development and deployment within their domains. The Order also established the first mandates for testing and disclosure for leading AI models ([see CED Backgrounder for detail](#)). This week the White House [released](#) a progress report, confirming the achievement of all actions designated for completion within 180 days of the Order. Those items include the National Institute for Standards and Technology releasing drafts for comment on [Generative AI](#), [Secure Software Development](#), and a [Plan for Global Engagement on AI Standards](#); new guidelines from CISA on [Mitigating AI Risks to Critical Infrastructure](#); and the Commerce Department's request for comment on [AI Accountability](#) related to AI audits and certifications.

In recent months, a number of bipartisan legislative proposals have also been introduced in Congress outlining possible elements of a US AI framework. The [Future of AI Innovation Act](#) would establish a Foundation Models Test Program within Commerce, codify the new US AI Safety Institute, and form an International AI Innovation and Standards Coalition. The proposed [Framework for Mitigating Extreme AI Risks](#) would establish a required evaluation and licensing scheme for frontier AI models, while it remains agnostic whether oversight would be through a new agency or pre-existing authorities. The [Bipartisan Framework for US AI Act](#) would similarly create a licensing regime and be administered by an Independent Oversight Body with authority to audit companies seeking licenses. The Bipartisan

Framework, sponsored by Senators Richard Blumenthal (D-CT) and Josh Hawley (R-MO), also clarifies that existing [Section 230](#) immunity for platforms publishing third-party content would not apply to AI companies. While no AI legislation has progressed in Congress, the growing number of bipartisan and bicameral efforts to outline frameworks indicates the potential for finding consensus.

European Union

The EU's [AI Act](#) will be the first comprehensive AI law globally after [final approval](#) by lawmakers in March. The law is on track to enter into force later this year, with the majority of provisions taking effect within two years. The Act will regulate frontier general-purpose AI models and classify AI applications according to their risk levels, with the high-risk tier having with new transparency obligations. Those requirements include registration of high-risk AI models in a centralized EU database and undergoing conformity tests monitored by the EU's new [AI Office](#). This will apply to AI developers, distributors, and employers with a presence in the EU market. The EU's framework relies heavily on its [General Data Protection Regulation](#) (GDPR), which went into force in 2018 and quickly became the [global standard](#) by applying to multinational companies with an EU presence. Europe could again set the [global standard](#) by leading the global regulatory push around AI with the first binding framework.

United Kingdom

From the outset, the UK has diverged from the EU's AI efforts, instead mirroring the US with a [pro-innovation approach](#) that first focuses on developing a framework for existing regulators to interpret and apply within their sector-specific domains. Select UK regulatory agencies were required to [publish](#) their AI annual strategic plans by April 30, similar to the early work of US agencies. The UK has prioritized initiatives outlined in its [National AI Strategy](#) from 2021, launching AI [research](#) programs, piloting a new [AI Standards Hub](#) to coordinate with global standardization, and [updating](#) cross-government standards for AI transparency. However, without introducing new laws or regulation, the framework currently relies on voluntary safety and transparency measures for developers of highly capable AI models. UK officials stated in a follow-up [response](#) in February that if AI capabilities continue to expand exponentially, binding measures could be produced if voluntary commitments were deemed insufficient.

China

After the EU, China is the second jurisdiction on the forefront of AI regulation, publishing its [Interim Measures for the Management of Generative Artificial Intelligence Services](#) in July before the measures took effect in August 2023 ([see CED Backgrounder for detail](#)). China's framework signaled plans to adopt a risk classification scheme — similar to the EU's AI Act — however, the guidelines do not yet specify what elements (e.g. risk types or industries) will factor into the classifications. Importantly, while the EU and Western governments have prioritized establishing guardrails or complete bans of the highest-risk AI use-cases, such as biometric surveillance, China's interim measures offer no comparable list of high-risk activities and how they will be handled. These new regulations are the latest in a series meant to address AI and algorithmic systems comprehensively, following the [Algorithm Recommendation Regulation](#) (effective March 2022) and its [Deep Synthesis Regulation](#) (effective January 2023), which establish new security assessments for AI tools that could influence public opinion.

Asia

Other Asian countries have largely taken a light-touch approach. The 10-country Association of Southeast Asian Nations (ASEAN), which includes Singapore, agreed to its [Guide on AI Governance and Ethics](#) in February. European officials had [lobbied](#) for the bloc to mirror the EU's comprehensive AI Act, including provisions on copyright and AI-generated content. Instead, the ASEAN guide followed the template of the US National Institute of Standards and Technology's (NIST) [AI Risk Management Framework](#) — a

voluntary set of best-practices. Some Critics have [commented](#) that the wide gap in ASEAN countries' digital capabilities and capacity of regulatory authorities provide different policy priorities among countries.

Singapore itself has produced voluntary AI governance frameworks, underscored by its [National AI Strategy 2.0](#) released in December. Singapore aims to be a global AI leader by focusing on bolstering its workforce and remaining attractive to foreign investment. The country's [Model AI Governance Framework](#) updated in 2020, and its companion [Implementation and Self-Assessment Guide](#), remain the foundation for voluntary standards. Indicating openness to global coordination, Singapore recently partnered with the US to build the [VerifyAI Initiative](#), which will act as a “crosswalk” between Singapore's AI framework and NIST's AI Risk Management Framework “focused on advancing shared principles and deepening information exchanges for safe, trustworthy, and responsible AI innovation.”

Japan also has promoted “agile” or non-binding governance, largely allowing the private sector to self-regulate. Its [AI Governance in Japan](#) white paper published in July 2021 remains the standard; it comprehensively describes Japan's AI regulatory policy, highlighting that “legally-binding horizontal requirements for AI systems are deemed unnecessary at the moment.”

Conclusion

The outlook for further international agreement on AI safety must now account for countries' domestic regulatory efforts, with differing priorities and tolerance for regulation. While the global community is coalescing around sets of shared principles, including on security and non-discrimination, national development of domestic AI policies exposes divergent approaches.

The US and UK remain a model for a focus on safety and lighter-touch regulation. Each has also undertaken similar initiatives, first prioritizing government uses of AI and clarifying the application of existing laws, while relying primarily on voluntary guidance for AI developers and users. Those still-developing policies will soon be confronted by binding frameworks established in the EU and China, the first and most comprehensive frameworks to apply to multinational companies.

While government AI initiatives are likely to make substantial progress in the coming year, further international agreement may prove difficult as countries turn their attention to more specific policy elements, including challenging questions such as copyright and market structure. As policymakers in the US and globally deliberate the structure of their national frameworks, the very different approach in the EU and China may have significant implications for US companies doing business in those jurisdictions. Truly global agreement on deeper principals for AI guardrails in more controversial areas such as data privacy, copyright, and surveillance, appears to be out of reach.

About the Authors

[Mitchell Barnes](#) is a Senior Economic Policy Analyst at the Committee for Economic Development, the public policy center of The Conference Board.

[John Gardner](#) is Vice President, Public Policy at the Committee for Economic Development, the public policy center of The Conference Board.

About The Conference Board

The Conference Board is the member-driven think tank that delivers Trusted Insights for What's Ahead™. Founded in 1916, we are a non-partisan, not-for-profit entity holding 501 (c) (3) tax-exempt status in the United States. www.ConferenceBoard.org

The Committee for Economic Development (CED) is the public policy center of The Conference Board. The nonprofit, nonpartisan, business-led policy center delivers trusted insights and reasoned solutions in the nation's interest. CED Trustees are chief executive officers and key executives of leading US companies who bring their unique experience to address today's pressing policy issues. Collectively, they represent 30+ industries and over 4 million employees. www.ConferenceBoard.org/us/committee-economic-development

© 2024 The Conference Board, Inc. All rights reserved.