International AI Policy Focuses on Safety

December 10, 2024

Jessica Renier, Managing Director, Digital Finance
Hannah Anderson, Policy Advisor, Digital Finance
Lokesh Bulchandani, Senior Research Analyst, Digital Finance
Evelyn Perfall, Program Assistant, Digital Finance

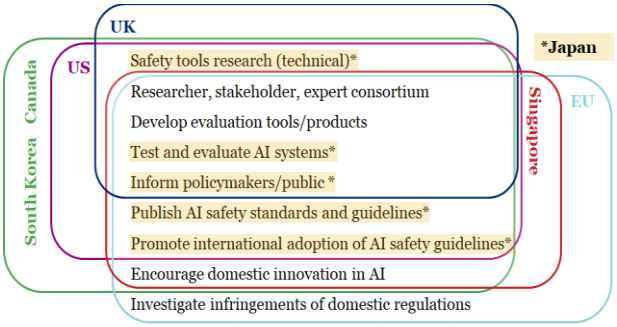
- Governments and international organizations have recognized the urgency of streamlining their approaches to AI safety, reflected in recent joint communiques, prompting tangible efforts at international collaboration.
- AI safety institutes (AISIs) are becoming a crucial mechanism for harmonizing safety standards and building consensus around the definition of safe AI; a newly established network of institutes will be crucial to watch.
- Alignment of international AI safety standards could ease the operational burden for firms navigating multiple jurisdictions, but transitioning to unified frameworks may introduce new complexities as firms grapple with new regulations.

SAFE COORDINATION ON AI Ensuring artificial intelligence's (AI) safe and responsible use and development has become a critical focus for multilateral, bilateral, and national organizations simultaneously. Initially, national authorities spearheaded efforts to regulate AI, but international collaboration and standards alignment are increasing in importance. This shift reflects the growing understanding that the challenges and opportunities posed by AI transcend borders, necessitating harmonized governance approaches. Global standard setters, and the G7 and G20, have emphasized this in recent statements. In a November report, the OECD named fostering international cooperation on AI development, deployment, and management as one of its ten policy priorities for managing AI risks. Managing potential risks to financial stability from widespread AI use will also require greater international cooperation and information sharing, according to the FSB's latest report on AI. Countries continue to commit to common principles for responsible AI use, such as in Europe's new AI Convention, the first binding treaty on AI safety. The needed alignment between all of these commitments is increasingly taking place through AI safety institutes (AISI). A newly launched Network of AISIs may model this desired cooperation. AI policies will also need to evolve at the same pace as the technology, complicating international coordination as multilateral cooperation is frequently a slow process. AISIs establish single actors dedicated to keeping pace with innovation and incorporating updates into international frameworks, if needed. Close engagement with these institutes will be vital for industry, as their importance looks to increase along with the complexity of guidelines governing AI use at a global scale.

A NETWORK FOR SAFETY Governments in multiple markets have independently developed AI safety measures tailored to their unique needs and regulatory frameworks. These policymakers are tasking AISIs with establishing harmonization between these policies. Set up at the national level, AISIs are specialized organizations that monitor the development of AI tools and their deployment, without a singular sector focus, and act as resource centers to developers and users to ensure AI tools are built in line with the standards set by national or multilateral bodies. Safety institutes also serve as hubs for technical evaluations, research, and dialogue between governments and industries. Major international policy gatherings, such as the second safety summit, the AI Seoul Summit in May 2024, have focused on AISI work, as we highlighted in our May 28 *Developments*. The ministers' statement at the Seoul Summit specifically highlighted the role of AISIs in fostering government-to-government collaboration, sharing risk management frameworks, and aligning safety standards across jurisdictions.

A subset of the 27 ministers signing onto the Seoul Declaration met again recently to form the International Network of AI Safety Institutes (INASI). Launched on November 21, 2024, in San Francisco, INASI unites AI safety efforts from ten governments—Australia, Canada, the EU, France, Japan, Kenya, Singapore, South Korea, the UK, and the U.S. Building on their agreed common goals from Seoul, INASI tasks member institutes with nine core functions. Multiple gatherings of policymakers and reviewing what each individual AISI has accomplished thus far suggests most Network members are actively performing their tasked functions, with some ambiguity as to their responsibility vis a vis existing national-level policy efforts, while some markets have yet to even begin (Chart).

Chart: INASI functions have not been operationalized by all markets equally



Source: CSIS, IIF. Australia, France, and Kenya have yet to establish AISIs tasked with these functions.

By performing these functions, policymakers plan for the INASI to achieve three vital objectives: 1) collaborate with the global scientific community to understand the risks and capabilities of advanced AI systems, sharing key findings throughout the network to better inform policy and safety standards; 2) develop testing frameworks through joint exercises and establish common best practices for evaluating

AI systems' safety and reliability; 3) promote inclusivity by engaging stakeholders from all regions, particularly developing nations, to ensure AI safety principles are globally accessible and applicable. For the financial services industry, INASI's work on AI-driven fraud and efforts to harmonize the regulatory landscape for AI deployers will be particularly noteworthy.

WORK UNDERWAY Network collaboration is already underway, demonstrating the capacity for some of the AISIs to fulfil these mandated functions. The U.S. and U.K. AI Safety Institutes recently released a joint technical report on pre-deployment testing of advanced AI systems. These AISIs brought together experts from both Institutes, and called upon leading scientists in their respective jurisdictions, to assess a to-be-released next generation large language model across four key domains: biological capabilities, cyber capabilities, software and AI development, and safeguard efficacy. This pre-deployment test revealed significant concerns about the ability of users to circumvent safeguards allowing the AISIs to recommend fixes and raise awareness of these risks before the model was released to the market. Pre-deployment testing efforts on this scale are crucial for financial services firms. If AISIs are identifying potential risks before AI systems are deployed, financial institutions can better assess the safety, reliability, and ethical implications of these technologies, ensuring they align with regulatory requirements and minimize exposure to emerging threats, before investing extensive resources. Additional AISI cooperation is likely to be announced soon, especially ahead of the February AI Action Summit in France, which will be the third in this safety summit series.

NATIONAL VARIATION AISIs are a new concept for many markets and much work is left to be done in their establishment. Though a participant in the INASI, and the EU's AI Act mandates their establishment, many members of the European Union are without an AISI. Major markets for AI development and deployment—France, Italy, and Germany—are in the process of setting up their AISIs. Further, each institute has varying degrees of authority and specialization, complicating their ability to take coordinated steps. The EU's new AI Office, created under the AI Act, is meant to work across these Institutes in an oversight capacity, and holds some safety functions of its own, with a focus on transparency. Japan, Singapore, South Korea (as of November 27), the UK, and the U.S. each have their own funded and functioning AISIs. Recent collaboration between UK and U.S. AISI's were discussed earlier, while Japan's AISI has focused on developing guides for evaluating safety at the individual user level and Singapore's AISI has focused on industry dialogues. South Korea's just-launched AISI's powers are still unclear.

A newer member of the INASI's active institutes club is the Canadian AI Safety Institute (CAISI). Launched in November 2024, CAISI is part of Canada's broader AI strategy, which includes a $2.4 billion investment in the Government's 2024 Budget, the Pan-Canadian Artificial Intelligence Strategy, the proposed Artificial Intelligence and Data Act, and the Voluntary Code of Conduct for Advanced Generative AI Systems. Collectively, this set of policies and investments will support future Canadian leadership in AI, while setting out clear limits on AI's acceptable use and appropriate protections for CAISI to coordinate across industries, and with other INASI members. This new AISI represents a much more empowered construction than its European counterparts, raising questions about just how coordinated INASI's work can be, especially as Kenya and Australia have yet to create their own AISIs.

NOT JUST SAFETY POLICIES Countries are pressing ahead with their own AI policies, adding additional considerations for AISIs to coordinate and work into existing frameworks. Argentina, Hong Kong, and New York State have each recently issued new guidance on risk management and acceptable uses for AI. In the EU, the European Commission (EC) is seeking input on defining unacceptable high-risk use cases for AI systems through a consultation. The IIF is currently seeking member feedback on its response letter and expects this consultation to inform guidelines to be published in early 2025. Meanwhile, the European AI Office has published the first draft of a General-Purpose AI Code of Practice, which outlines expectations for the conduct of general-purpose AI model providers. If this proliferation of national-level AI policies continues at this rate, multinational firms may increasingly struggle to understand and fulfil their compliance obligations. There may also be more opportunities for conflicting regulations, potentially limiting efficiency gains from widespread AI deployment.

CONCLUSION While every jurisdiction may develop its own framework for AI safety, effective global governance of AI requires interoperability in these frameworks. Safety institutes will play a pivotal role in addressing this challenge by conducting evaluations, fostering foundational research, and facilitating information exchange. In the long term, aligned global standards promise to reduce regulatory fragmentation, streamline compliance requirements, and enable more seamless operations across multiple jurisdictions. Getting to this alignment may be difficult, as firms still must confront the matrix of national and international AI usage policies, especially in markets with nascent AI rules. The growing network of AISIs will be an important venue for engagement and advocacy going forward.