The International Network of AI Safety Institutes

Mission Statement

The AI safety institutes and government mandated offices that facilitate AI safety and evaluation from Australia, Canada, the European Commission, France, Japan, Kenya, the Republic of Korea, Singapore, the United Kingdom, and the United States, have gathered today in San Francisco on November 21, 2024 to launch the International Network of AI Safety Institutes. Affirming and building on the *Seoul Statement of Intent toward International Cooperation on AI Safety Science* released at the AI Seoul Summit on May 21, 2024, this Network is intended to catalyze a new phase of international cooperation on AI safety.

Our institutes and offices are technical organisations that aim to advance AI safety, help governments and society understand the risks posed by advanced AI systems, and suggest solutions to address those risks in order to minimize harm. Beyond mitigating risks, these institutes and offices play a crucial role in guiding the responsible development and deployment of AI systems. AI presents enormous opportunities – the ability to serve societal needs and transform and enhance human wellbeing, peace, and prosperity – as well as potential global risks. International cooperation to promote AI safety, security, inclusivity, and trust is essential to address those risks, drive responsible innovation, and expand access to the benefits of AI worldwide.

The International Network of AI Safety Institutes is intended to be a forum that brings together technical expertise from around the world. Recognizing the importance of cultural and linguistic diversity, we aim to facilitate a common technical understanding of AI safety risks and mitigations based upon the work of our institutes and of the broader scientific community that will support international development and the adoption of interoperable principles and best practices. We also intend to encourage a general understanding of and approach to AI safety globally, that will enable the benefits of AI innovation to be shared amongst countries at all stages of development. The leadership of our respective AI safety institutes and offices have therefore come together to collaborate as Network Members to help drive technical alignment on AI safety research, testing, and guidance.

As a first step, we intend to prioritize our collective efforts, in line with our mandates, in the following four areas crucial to international AI safety:

- **Research:** We plan, together with the scientific community, to advance research on risks and capabilities of advanced AI systems as well as to share the most relevant results, as appropriate, from research that advances the science of AI safety.
- **Testing:** We plan to work towards building common best practices for testing advanced AI systems. This work may include conducting joint testing exercises and sharing results from domestic evaluations, as appropriate.
- **Guidance:** We plan to facilitate shared approaches such as interpreting tests of advanced systems, where appropriate.
- **Inclusion:** We plan to actively engage countries, partners, and stakeholders in all regions of the world and at all levels of development by sharing information and technical tools in an accessible and collaborative manner, where appropriate. We hope, through these actions, to increase the capacity for a diverse range of actors to participate in the science and practice of AI safety. Through this Network, we are dedicated to collaborating broadly with partners to ensure that safe, secure, and trustworthy AI benefits all of humanity.