Annex

Seoul Statement of Intent toward International Cooperation on AI Safety Science

1. Gathered at the AI Seoul Summit on 21st May 2024, and following on from the AI Safety Summit in Bletchley Park on 2nd November 2023 and acknowledging the Safety Testing Chair's Statement of Session Outcomes from the Bletchley Leaders' Session, world leaders representing Australia, Canada, the European Union, France, Germany, Italy, Japan, the Republic of Korea, the Republic of Singapore, the United Kingdom, and the United States of America affirm the importance of international coordination and collaboration, based in openness, transparency, and reciprocity, to advance the science of AI safety. We affirm that safety is a key element in furtherance of responsible AI innovation.

2. We commend the collective work to create or expand public and/or government-backed institutions, including AI Safety Institutes, that facilitate AI safety research, testing, and/or developing guidance to advance AI safety for commercially and publicly available AI systems.

   2.1 We acknowledge the need for a reliable, interdisciplinary, and reproducible body of evidence to inform policy efforts related to AI safety. We recognize the role of scientific inquiry and the benefits of international coordination for the advancement of such inquiry, so that ultimately the benefits of AI development and deployment are shared equitably around the globe.

   2.2 We affirm our intention to leverage and promote common scientific understandings through assessments such as the *International AI Safety Report,* to guide and align our respective policies, where appropriate, and to enable safe, secure, and trustworthy AI innovation, in line with our governance frameworks.

   2.3 We express our shared intent to take steps toward fostering common international scientific understanding on aspects of AI safety, including by endeavoring to promote complementarity and interoperability in our technical methodologies and overall approaches.

2.4 These steps may include taking advantage of existing initiatives; the mutual strengthening of research, testing, and guidance capacities; the sharing of information about models, including their capabilities, limitations, and risks as appropriate; the monitoring of AI harms and safety incidents; the exchange or joint creation of evaluations, data sets and associated criteria, where appropriate; the establishment of shared technical resources for purposes of advancing the science of AI safety; and the promotion of appropriate research security practices in the field.

2.5 We intend to coordinate our efforts to maximize efficiency, define priorities, report progress, enhance our outputs' scientific rigor and robustness, promote the development and adoption of international standards, and accelerate the advancement of evidence-based approaches to AI safety.

3. We articulate our shared ambition to develop an international network among key partners to accelerate the advancement of the science of AI safety. We look forward to close future collaboration, dialogue, and partnership on these and related endeavors.