SEOUL MINISTERIAL STATEMENT FOR ADVANCING AI SAFETY, INNOVATION AND INCLUSIVITY, AI SEOUL SUMMIT, 22nd MAY 2024

Preamble

P1. We, the Ministers of Australia, Canada, Chile, France, Germany, India, Indonesia, Israel, Italy, Japan, Kenya, Mexico, the Netherlands, Nigeria, New Zealand, the Philippines, the Republic of Korea, Rwanda, the Kingdom of Saudi Arabia, the Republic of Singapore, Spain, Switzerland, Türkiye, Ukraine, the United Arab Emirates, the United Kingdom, the United States of America, and the representative of the European Union, based on the discussion at the AI Seoul Summit Minister's Session on 22nd May 2024 on the approaches to AI governance to promote safe, secure and trustworthy AI and sustainable AI development, hereby affirm the need for collaborative international approaches to respond to rapid advancements in AI technologies and their impact on our societies and economies.

P2. We acknowledge the achievements initiated at the inaugural AI Safety Summit in Bletchley Park. Building upon the three inter-related priorities of safety, innovation, and inclusivity discussed on day 1 of the AI Seoul Summit, we now seek to focus on actionable items to implement them. Acknowledging our different domestic capacities, we affirm our shared intent to take meaningful steps to unlock the benefits of AI for all while addressing its risks.

Safety

1.1 It is imperative to guard against the full spectrum of AI risks, including risks posed by the deployment and use of current and frontier AI models or systems and those that may be designed, developed, deployed and used in future. Principles for AI safety and security include transparency, interpretability and explainability; privacy and accountability; meaningful human oversight and effective data management and protection. We encourage all relevant actors, including organizations developing and deploying current and frontier AI, to promote accountability and transparency throughout the AI lifecycle by seeking to assess, prevent, mitigate and remedy adverse impacts which may emerge. We further encourage all relevant actors to foster an enabling environment in which AI is designed, developed, deployed and used in a safe, secure and trustworthy manner, for the good of all and in line with applicable domestic and international frameworks.

1.2 We recognize our role to establish frameworks for managing risks posed by the design, development, deployment and use of commercially or publicly available frontier AI models or systems in our respective jurisdictions. We recognize our increasing role in promoting credible external evaluations for frontier AI models or systems developed in our jurisdictions, where those models or systems could pose severe risks. We further acknowledge our role in partnership with the private sector, civil society, academia and the international community in identifying thresholds at which the risks posed by the design, development, deployment and use of frontier AI models or systems would be severe without appropriate mitigations. Criteria for assessing the risks posed by frontier AI models or systems may include consideration of capabilities, limitations and propensities, implemented safeguards, including robustness against

malicious adversarial attacks and manipulation, foreseeable uses and misuses, deployment contexts, including the broader system into which an AI model may be integrated, reach, and other relevant risk factors.

1.3 Assessing the risk posed by the design, development, deployment and use of frontier AI models or systems may involve defining and measuring model or system capabilities that could pose severe risks, in context and without appropriate mitigations. We recognize that such severe risks could be posed by the potential model or system capability to meaningfully assist non-state actors in advancing the development, production, acquisition or use of chemical or biological weapons, as well as their means of delivery. We affirm the continuing importance of acting consistently with relevant international law, such as the Chemical Weapons Convention and Biological and Toxin Weapons Convention, UN Security Council Resolution 1540, and international human rights law, in accordance with each state's obligations. We stress the importance of multilateral discussion to promote AI safety and security.

1.4 We further recognize that such severe risks could be posed by the potential model or system capability or propensity to evade human oversight, including through safeguard circumvention, manipulation and deception, or autonomous replication and adaptation conducted without explicit human approval or permission. We note the importance of gathering further empirical data with regard to the risks from frontier AI models or systems with highly advanced agentic capabilities, at the same time as we acknowledge the necessity of preventing the misuse or misalignment of such models or systems, including by working with organizations developing and deploying frontier AI to implement appropriate safeguards, such as the capacity for meaningful human oversight.

1.5 We acknowledge the importance of constructive dialogue with developers to address the risks of frontier AI models or systems, reaffirming the particular responsibility of developers for the safety of these systems. We further recognize the pressing need to take into consideration safety and security throughout the AI lifecycle.

1.6 We affirm the unique role of AI safety institutes and other relevant institutions to enhance international cooperation on AI risk management and increase global understanding in the realm of AI safety and security. Through our AI safety institutes or other relevant institutions, we plan to share best practices and evaluation datasets, as appropriate, and collaborate in establishing safety testing guidelines. We aim towards interoperability across AI safety activity, including by building partnerships between AI safety institutes and other relevant institutions, recognizing at the same time the need for testing methodologies considering cultural and linguistic diversity across the globe.

Innovation

2.1 We recognize the importance of governance approaches that foster innovation and the development of AI industry ecosystems with the goal of maximizing the potential benefits of AI for our economies and societies. We further recognize the role of governments is not only to

prioritize financial investment, R&D, and workforce development for AI innovation, but also to consider governance frameworks which include legal and institutional frameworks, including personal data, copyright and other intellectual property protections for the safe, secure and trustworthy development and deployment of AI.

2.2 We recognize the transformative benefits of AI for the public sector, including in areas such as administration, welfare, education and healthcare. These benefits include using AI efficiently and effectively through accessible digital services and automated procedures that enhance citizen experience in accessing public services. Furthermore, we intend to support the adoption of AI in key industrial sectors like manufacturing, logistics, and finance to revolutionize productivity, reduce the burden on employees while protecting rights and safety and unlock new avenues for value creation.

2.3 We are committed in particular, to supporting an environment conducive to AI-driven innovation by facilitating access to AI-related resources in particular for SMEs, startups, academia, universities, and even individuals, while respecting and safeguarding intellectual property rights. Also, we are committed to enhancing the availability of AI-related resources to empower researchers to leverage AI in their respective fields of study and to facilitate the responsible utilization of AI as a tool for enriching individual creative endeavors.

2.4 We recognize the importance of sustainability and resilience in the ecosystem for AI innovation. In this regard, we encourage AI developers and deployers to take into consideration their potential environmental footprint such as energy and resource consumption. We welcome collaborative efforts to explore measures on how our workforce can be upskilled and reskilled to be confident users and developers of AI to enhance innovation and productivity. Furthermore, we encourage efforts by companies to promote the development and use of resource-efficient AI models or systems and inputs such as applying low-power AI chips and operating environmentally friendly data centers throughout AI development and services.

Inclusivity

3.1 In our efforts to foster an inclusive digital transformation, we recognize that the benefits of AI should be shared equitably. We seek to promote our shared vision to leverage the benefits of AI for all, including vulnerable groups. We intend to work together to promote the inclusive development of AI systems and the utilization of safe, secure and trustworthy AI technologies in order to foster our shared values and mutual trust. We recognize the potential of AI for the benefit of all, especially in protecting human rights and fundamental freedoms, strengthening social safety nets, as well as ensuring safety from various risks including disasters and accidents.

3.2 In furtherance of our shared goal to inclusivity, we are committed to promoting AI education including through capacity-building related to AI systems and through increased digital literacy, contributing to bridging AI and digital divides between and within countries. We recognize the need to strengthen international cooperation in joint research and talent development, including with developing countries to enhance their capabilities in AI design, development and utilization. We seek to ensure socio-cultural and linguistic diversity is reflected and promoted in the AI

lifecycle of design, development, deployment, and use.

3.3 We are committed to supporting and promoting advancements in AI technologies, recognizing the potential to provide significant advances to resolve the world's greatest challenges such as climate change, global health, food and energy security and education. We further seek to foster inclusive governance approaches by encouraging the participation of developing countries in joint efforts and discussions aimed at accelerating progress toward achieving the Sustainable Development Goals and promoting global common interests and developments.

Way Forward

W1. We commend the efforts undertaken by the Republic of Korea and the United Kingdom as the two co-chairs of the AI Seoul Summit on the agendas of safety, and sustainability and resilience.

W2. We note the publication of the independent interim *International Scientific Report on the Safety of Advanced AI* and its work to facilitate a shared evidence-based understanding of the risks associated with frontier AI. We resolve to work together to advance future evidence-based reports on AI risk, and look forward to the final publication of the *International Scientific Report on the Safety of Advanced AI* ahead of the AI Action Summit in France.

W3. We acknowledge the need to advance the science of AI safety and gather more empirical data with regard to certain risks, at the same time as we recognise the need to translate our collective understanding into empirically grounded, proactive measures with regard to capabilities that could result in severe risks. We plan to collaborate with the private sector, civil society and academia, to identify thresholds at which the level of risk posed by the design, development, deployment and use of frontier AI models or systems would be severe absent appropriate mitigations, and to define frontier AI model or system capabilities that could pose severe risks, with the ambition of developing proposals for consideration in advance of the AI Action Summit in France.

W4. We reaffirm our shared intent to guide the design, development, deployment, and use of AI in a manner which harnesses its benefits for good. With the recognition that safety, innovation, and inclusivity are inter-related goals, we look forward to continuing our collaboration to advance discussions on AI governance and promote safe, secure and trustworthy AI for the good of all.