ALERT MEMORANDUM

G7 Leaders Publish AI Code of Conduct: A Common Thread in the Patchwork of Emerging AI Regulations Globally?

November 1, 2023

On October 30, 2023, the G7 Leaders published a Statement on the Hiroshima Artificial Intelligence ("**AI**") Process (the "**Statement**").[1] This follows the G7 Summit in May, where the leaders agreed on the need to address the risks arising from rapidly evolving AI technologies. The Statement was accompanied by the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (the "**Code of Conduct**")[2] and the Hiroshima Process International Guiding Principles for Advanced AI Systems (the "**Guiding Principles**").[3]

The Code of Conduct sets out voluntary guidance for private sector and other organizations developing and using advanced AI systems. The Code of Conduct does not define conclusively an "advanced AI system" but contemplates that advanced foundation models and generative AI systems will be covered. The Code of Conduct is arranged around the Guiding Principles, and aims to promote safe, secure, and trustworthy AI. In particular, it emphasizes the importance of adopting a risk-based approach to implementation of certain actions.

This alert memorandum summarizes the background to this initiative, certain key points of the Code of Conduct and Guiding Principles, and possible next steps.

If you have any questions concerning this memorandum, please reach out to your regular firm contact or the following authors

LONDON

Henry Mostyn
+44 20 7614 2241
hmostyn@cgsh.com

Gareth Kristensen
+44 20 7614 2381
gkristensen@cgsh.com

Ferdisha Snagg
+44 20 7614 2251
fsnagg@cgsh.com

Prudence Buckland
+44 20 7614 2317
pbuckland@cgsh.com

Anders Jay
+1 650 815 4155
ajay@cgsh.com

Andreas Wildner
+44 20 7614 2248
awildner@cgsh.com

[1] The G7 Hiroshima AI Process Statement is accessible here.

[2] The G7 Hiroshima AI Process Code of Conduct is accessible here.

[3] The G7 Hiroshima AI Process Guiding Principles are accessible here.

Context

On May 19, 2023, the G7 Leaders convened in Hiroshima for their annual Summit. One of the outcomes of that summit was the establishment of the Hiroshima AI Process. This is effectively a G7 working group tasked with taking stock of the opportunities and challenges flowing from AI, and discussing topics such as governance, intellectual property and data privacy protections, responsible utilization of AI technologies, promoting transparency, and responding to information manipulation and disinformation (particularly in the context of generative AI).[4]

The Hiroshima AI Process seeks to complement ongoing discussions within a number of international forums, including the Organization for Economic Co-operation and Development (the "**OECD**") and the Global Partnership on Artificial Intelligence as well as the EU-U.S. Trade and Technology Council and the EU's Digital Partnerships with Japan, Korea and Singapore.

AI is also an increasingly prominent item on G7 jurisdictions' domestic policy-making agendas. For example, in the U.S., several leading AI organizations have agreed voluntary commitments on safety, security, and transparency with the government,[5] and on October 30, President Biden issued an Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.[6] The EU has also proposed the AI Act: a broad regulatory framework for AI with different requirements dependent on the risk associated with certain uses of the technology.[7] The UK Government is also hosting an AI Safety Summit on November 1 and 2, bringing together international governments, leading AI companies, civil society groups and experts in research, to consider the risks of AI, especially at the frontier of development and discuss how they can be mitigated through internationally coordinated action.[8] On November 1, the governments of several countries attending the AI Safety Summit 2023 signed the Bletchley Declaration, affirming their commitment to international cooperation with a view to identifying AI safety risks and the impact of AI on society, and building respective risk-based policies across the various countries.[9] This comes further to the UK Competition and Markets Authority's initial review of AI foundation models, which looked at the risks and opportunities AI may bring from a competition and consumer protection standpoint.[10]

Code of Conduct and Guiding Principles

The Statement emphasizes the opportunities that advanced AI systems may bring while also highlighting the risks and challenges posed by such technology, in particular possible systemic risks.

The Code of Conduct sets out steps organizations are expected to take with respect to development and use of such AI technologies. It does so through incorporating, and elaborating on, the eleven principles set out in the Guiding Principles.

— **Take appropriate measures to identify, evaluate and mitigate risks across the lifecycle of advanced AI systems, including prior to and throughout deployment/placement on the market**. This should be done through a combination of methods for evaluation and testing and other risk mitigation measures. Testing should take place in secure environments, before deployment on the market. AI developers should ensure traceability (e.g., in relation to datasets, processes and decisions made during system development), and should document measures and

[4] See G7 Hiroshima Leaders' Communiqué of 20 May 2023, accessible here.

[5] For further information on the voluntary commitments from leading AI companies to manage the risks posed by AI, please see the US Government's announcement here.

[6] For further information on President Biden's Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, please see the US Government's announcement here.

[7] The European Commission's Proposal for an AI Act is accessible here.

[8] For further information on the UK's AI Safety Summit 2023, see here. Calls for international panels on AI safety have been put forward on various occasions, including by prominent figures from the industry (see, e.g., here).

[9] The Bletchley Declaration is accessible here.

[10] See CMA, AI Foundational Models Initial Report, 18 September 2023 (available here).

keep technical documentation up-to-date. The Code of Conduct lists a number of risks that organizations should devote attention to, including offensive cyber capabilities, risks related to weapons development/acquisition/use, or societal risks.

— **Identify and mitigate vulnerabilities, incidents and patterns of misuse after deployment/ placement on the market**. Commensurate to the level of risk posed by an AI system, organisations should monitor for, and implement mechanisms to report, vulnerabilities, incidents, emerging risks and technology misuse. This might include, for example, facilitating third-party and user discovery and reporting of issues and vulnerabilities.

— **Publicly report advanced AI systems' capabilities, limitations and domains for appropriate and inappropriate use to support transparency and accountability**. This should include publishing transparency reports, instructions for use and relevant technical documentation. These should contain information on the evaluations conducted and the results, on capacities and limitations of an AI model or system, and a discussion and assessment of the resultant effects and risks to safety and society. Reporting should be kept up-to-date, be sufficiently clear and understandable, and be supported by robust documentation processes.

— **Responsible information sharing and incident reporting among organizations developing advanced AI systems.** This may include evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards and other relevant documentation and transparency measures. Organizations should collaborate to develop, advance and adopt shared standards, tools, mechanisms and best practices for ensuring safety, security and trustworthiness of AI systems. In complying with this principle, organizations will need to carefully observe antitrust safeguards.

— **Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach.** Organisation should put in place appropriate organisational mechanisms to develop, disclose and implement risk management and governance policies, where feasible. This includes disclosing where appropriate privacy policies, user prompts and advanced AI system outputs. Policies should be developed in accordance with a risk-based approach, and be regularly updated.

— **Implement robust security controls, including physical security, cybersecurity and insider threat safeguards.** These may involve securing model weights, algorithms, servers and datasets through appropriate operational security measures and access controls, and implementing policies to address the same. Organisations should also consider establishing an insider threat detection program to protect key intellectual property and trade secrets.

— **Develop and deploy reliable mechanisms to enable users to identify AI-generated content / understand when they are interacting with an AI system**. This may include authentication and provenance mechanisms where feasible (e.g., to include an identifier of the service or model that created relevant content). Organizations should also implement mechanisms such as labelling or disclaimers to enable users to understand when they are interacting with AI systems.

— **Prioritise research to advance AI safety, security and trustworthiness, address key risks and develop mitigation tools.** This may involve conducting, investing in, and collaborating on research on key aspects (e.g., avoidance of harmful bias or information manipulation, or safeguarding IP rights and privacy). Mitigation tools should be developed to proactively manage risks of advanced AI systems, including environmental and climate impacts. Organizations are encouraged share research and best practices on risk mitigation.

— **Prioritize the development of advanced AI systems to address the world's greatest challenges.** Organisations are encouraged to

develop AI technologies to support progress on the UN Sustainable Development Goals, and help addressing challenges such as the climate crisis, global health and education. Organizations should support digital literacy initiatives to enable the wider public to benefit from the use of advanced AI systems.

— **Advance the development and adoption of interoperable international technical standards and best practices.** Examples of areas for standardization include watermarking, testing methodologies, content authentication and provenance mechanisms, cybersecurity policies, public reporting.

— **Implement appropriate data input measures and protections for personal data and intellectual property.** Organizations should take appropriate measures (e.g. transparency measures) to manage data quality and to mitigate against harmful biases. Moreover, organisations should implement measures to protect confidential or sensitive data, including with respect to the training, testing and fine-tuning of models. The Code of Conduct does not specify exactly how this should be done, and it is not clear how firms are expect to comply in practice. Organizations should also implement appropriate safeguards to respect privacy and IP rights.

Next Steps

The Statement notes that the Code of Conduct and Guiding Principles will be reviewed and updated as necessary, including through ongoing inclusive multistakeholder consultation.

Importantly, in addition the G7 Leaders instructed relevant ministers to develop, by the end of this year, a 'Hiroshima AI Process Comprehensive Policy Framework', including project-based cooperation with the OECD and Global Partnership on Artificial Intelligence, and a work plan for further advancing the Hiroshima AI Process.

It is not clear how the Code of Conduct and Guiding Principles will supplement the existing and emerging regulatory requirements applicable to development and use of AI in G7 countries in practice. In the EU, for example, the proposed AI Act, the proposed AI Liability Directive,[11] and the Digital Markets Act[12] will already subject actors in the AI value chain to various requirements, restrictions, and potential liabilities in respect of the AI technologies they may seek to develop and/or deploy. In the UK, where the Government has recommended context-specific guidance rather than uniform legislation, regulators (such as the UK Competition and Markets Authority, Financial Conduct Authority and Information Commissioner's Office) may draw on the Hiroshima materials in considering how to apply existing rules to AI related issues.

There are other areas where binding regulation and voluntary codes of conduct are being developed in parallel, for example, in the area of ESG rating providers in the UK.[13] The perceived advantages of voluntary codes of conduct in this respect may be that such measures can be used to address issues more quickly than binding regulation, and that market participants' experiences in adopting such measures can be taken into account when creating binding rules.

However, with AI regulation in G7 countries developing at pace and growing regulatory scrutiny of such technology, it is critical that co-legislators and rule-makers take care to ensure legal certainty with respect to how any binding (or non-binding) measures will apply – particularly in areas of overlap between different sets of rules. It will also be crucial to ensure that these measures do not conflate foundation models with the AI systems that may integrate such models, and that such measures account for the role of different participants in the AI value chain and the purposes for

[11] The European Commission Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (the "**AI Liability Directive**") is accessible here.

[12] For a recent overview over the Digital Markets Act, please see here.

[13] Please see here for an overview of the developments regarding a UK regulation for ESG rating providers, and here for an analysis of the proposed voluntary code of conduct (the latter article is also available upon request).

which an AI system is deployed and used. This is consistent with the risk-based approach outlined in the Code of Conduct, and equivalent concepts in other regulatory regimes such as the EU's proposed AI Act.

…

CLEARY GOTTLIEB