Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems

On the basis of the International Guiding Principles for Organizations Developing Advanced AI systems, the International Code of Conduct for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide and will provide voluntary guidance for actions by organizations developing the most advanced AI systems, including the most advanced foundation models and generative AI systems (henceforth "advanced AI systems").

Organizations should follow these actions in line with a risk-based approach.

Organizations that may endorse this Code of Conduct may include, among others, entities from academia, civil society, the private sector, and/or the public sector.

This non-exhaustive list of actions is discussed and elaborated as a living document to build on the existing OECD AI Principles in response to the recent developments in advanced AI systems and is meant to help seize the benefits and address the risks and challenges brought by these technologies. Organizations should apply these actions to all stages of the lifecycle to cover, when and as applicable, the design, development, deployment and use of advanced AI systems.

This document will be reviewed and updated as necessary, including through ongoing inclusive multistakeholder consultations, in order to ensure it remains fit for purpose and responsive to this rapidly evolving technology.

Different jurisdictions may take their own unique approaches to implementing these actions in different ways.

We call on organizations in consultation with other relevant stakeholders to follow these actions, in line with a risk-based approach, while governments develop more enduring and/or detailed governance and regulatory approaches. We also commit to develop proposals, in consultation with the OECD, GPAI and other stakeholders, to introduce monitoring tools and mechanisms to help organizations stay accountable for the implementation of these actions. We encourage organizations to support the development of effective monitoring mechanisms, which we may explore to develop, by contributing best practices.

In addition, we encourage organizations to set up internal AI governance structures and policies, including self-assessment mechanisms, to facilitate a responsible and accountable approach to implementation of these actions and in AI development.

While harnessing the opportunities of innovation, organizations should respect the rule of law, human rights, due process, diversity, fairness and non-discrimination, democracy, and human-centricity, in the design, development and deployment of advanced AI systems.

Organizations should not develop or deploy advanced AI systems in ways that undermine democratic values, are particularly harmful to individuals or communities, facilitate terrorism, promote criminal misuse, or pose substantial risks to safety, security and human rights, and are thus not acceptable.

States must abide by their obligations under international human rights law to ensure that human rights are fully respected and protected, while private sector activities should be in line with international frameworks such as the United Nations Guiding Principles on Business and Human Rights and the OECD Guidelines for Multinational Enterprises.

Specifically, we call on organizations to abide by the following actions, in a manner that is commensurate to the risks:

1 Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.

This includes employing diverse internal and independent external testing measures, through a combination of methods for evaluations, such as red-teaming, and implementing appropriate mitigation to address identified risks and vulnerabilities. Testing and mitigation measures, should, for example, seek to ensure the trustworthiness, safety and security of systems throughout their entire lifecycle so that they do not pose unreasonable risks. In support of such testing, developers should seek to enable traceability, in relation to datasets, processes, and decisions made during system development. These measures should be documented and supported by regularly updated technical documentation.

This testing should take place in secure environments and be performed at several checkpoints throughout the AI lifecycle in particular before deployment and placement on the market to identify risks and vulnerabilities, and to inform action to address the identified AI risks to security, safety and societal and other risks, whether accidental or intentional. In

designing and implementing testing measures, organizations commit to devote attention to the following risks as appropriate:

> Chemical, biological, radiological, and nuclear risks, such as the ways in which advanced AI systems can lower barriers to entry, including for non-state actors, for weapons development, design acquisition, or use.

> Offensive cyber capabilities, such as the ways in which systems can enable vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system.

> Risks to health and/or Safety, including the effects of system interaction and tool use, including for example the capacity to control physical systems and interfere with critical infrastructure.

> Risks from models of making copies of themselves or “self-replicating” or training other models.

> Societal risks, as well as risks to individuals and communities such as the ways in which advanced AI systems or models can give rise to harmful bias and discrimination or lead to violation of applicable legal frameworks, including on privacy and data protection.

> Threats to democratic values and human rights, including the facilitation of disinformation or harming privacy.

> Risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.

Organizations commit to work in collaboration with relevant actors across sectors, to assess and adopt mitigation measures to address these risks, in particular systemic risks.

Organizations making these commitments should also endeavor to advance research and investment on the security, safety, bias and disinformation, fairness, explainability and interpretability, and transparency of advanced AI systems and on increasing robustness and trustworthiness of advanced AI systems against misuse.

2 ***Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.***

Organizations should use, as and when appropriate commensurate to the level of risk, AI systems as intended and monitor for vulnerabilities, incidents, emerging risks and misuse after deployment, and take appropriate action to address these. Organizations are encouraged to consider, for example, facilitating third-party and user discovery and reporting of issues and vulnerabilities after deployment such as through bounty systems, contests, or prizes to incentivize the responsible disclosure of weaknesses. Organizations are further encouraged to maintain appropriate documentation of reported incidents and to mitigate the identified risks and vulnerabilities, in collaboration with other stakeholders. Mechanisms to report vulnerabilities, where appropriate, should be accessible to a diverse set of stakeholders.

3 ***Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.***

This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems.
These reports, instruction for use and relevant technical documentation, as appropriate as, should be kept up-to-date and should include, for example;

> Details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights,
> Capacities of a model/system and significant limitations in performance that have implications for the domains of appropriate use,
> Discussion and assessment of the model's or system's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness, and
> The results of red-teaming conducted to evaluate the model's/system's fitness for moving beyond the development stage.

Organizations should make the information in the transparency reports sufficiently clear and understandable to enable deployers and users as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately; and that transparency reporting should be supported and informed by robust documentation processes such as technical documentation and instructions for use.

4 ***Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia***

This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle.

Organizations should establish or join mechanisms to develop, advance, and adopt, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems.

This should also include ensuring appropriate and relevant documentation and transparency across the AI lifecycle in particular for advanced AI systems that cause significant risks to safety and society.

Organizations should collaborate with other organizations across the AI lifecycle to share and report relevant information to the public with a view to advancing safety, security and trustworthiness of advanced AI systems. Organizations should also collaborate and share the aforementioned information with relevant public authorities, as appropriate.

Such reporting should safeguard intellectual property rights.

5 ***Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures.***

Organizations should put in place appropriate organizational mechanisms to develop, disclose and implement risk management and governance policies, including for example accountability and governance processes to identify, assess, prevent, and address risks, where feasible throughout the AI lifecycle.

This includes disclosing where appropriate privacy policies, including for personal data, user prompts and advanced AI system outputs. Organizations are expected to establish and disclose their AI governance policies and organizational mechanisms to implement these policies in accordance with a risk based approach. This should include accountability and

governance processes to evaluate and mitigate risks, where feasible throughout the AI lifecycle.

The risk management policies should be developed in accordance with a risk based approach and apply a risk management framework across the AI lifecycle as appropriate and relevant, to address the range of risks associated with AI systems, and policies should also be regularly updated.

Organizations should establish policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices

6 ***Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.***

These may include securing model weights and, algorithms, servers, and datasets, such as through operational security measures for information security and appropriate cyber/physical access controls.

This also includes performing an assessment of cybersecurity risks and implementing cybersecurity policies and adequate technical and institutional solutions to ensure that the cybersecurity of advanced AI systems is appropriate to the relevant circumstances and the risks involved. Organizations should also have in place measures to require storing and working with the model weights of advanced AI systems in an appropriately secure environment with limited access to reduce both the risk of unsanctioned release and the risk of unauthorized access. This includes a commitment to have in place a vulnerability management process and to regularly review security measures to ensure they are maintained to a high standard and remain suitable to address risks.

This further includes establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets, for example, by limiting access to proprietary and unreleased model weights.

7 ***Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content***

This includes, where appropriate and technically feasible, content authentication and

provenance mechanisms for content created with an organization's advanced AI system. The provenance data should include an identifier of the service or model that created the content, but need not include user information. Organizations should also endeavor to develop tools or APIs to allow users to determine if particular content was created with their advanced AI system, such as via watermarks. Organizations should collaborate and invest in research, as appropriate, to advance the state of the field.

Organizations are further encouraged to implement other mechanisms such as labeling or disclaimers to enable users, where possible and appropriate, to know when they are interacting with an AI system.

8 ***Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.***

This includes conducting, collaborating on and investing in research that supports the advancement of AI safety, security, and trust, and addressing key risks, as well as investing in developing appropriate mitigation tools.

Organizations commit to conducting, collaborating on and investing in research that supports the advancement of AI safety, security, trustworthiness and addressing key risks, such as prioritizing research on upholding democratic values, respecting human rights, protecting children and vulnerable groups, safeguarding intellectual property rights and privacy, and avoiding harmful bias, mis- and disinformation, and information manipulation. Organizations also commit to invest in developing appropriate mitigation tools, and work to proactively manage the risks of advanced AI systems, including environmental and climate impacts, so that their benefits can be realized.

Organizations are encouraged to share research and best practices on risk mitigation.

9 ***Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education***

These efforts are undertaken in support of progress on the United Nations Sustainable Development Goals, and to encourage AI development for global benefit.

Organizations should prioritize responsible stewardship of trustworthy and human-centric AI and also support digital literacy initiatives that promote the education and training of the

public, including students and workers, to enable them to benefit from the use of advanced AI systems, and to help individuals and communities better understand the nature, capabilities, limitations, and impact of these technologies. Organizations should work with civil society and community groups to identify priority challenges and develop innovative solutions to address the world's greatest challenges.

10 *Advance the development of and, where appropriate, adoption of international technical standards*

Organizations are encouraged to contribute to the development and, where appropriate, use of international technical standards and best practices, including for watermarking, and working with Standards Development Organizations (SDOs), also when developing organizations' testing methodologies, content authentication and provenance mechanisms, cybersecurity policies, public reporting, and other measures. In particular, organizations also are encouraged to work to develop interoperable international technical standards and frameworks to help users distinguish content generated by AI from non-AI generated content.

11 *Implement appropriate data input measures and protections for personal data and intellectual property*

Organizations are encouraged to take appropriate measures to manage data quality, including training data and data collection, to mitigate against harmful biases.

Appropriate measures could include transparency, privacy-preserving training techniques, and/or testing and fine-tuning to ensure that systems do not divulge confidential or sensitive data.

Organizations are encouraged to implement appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright-protected content.

Organizations should also comply with applicable legal frameworks.