International Draft Guiding Principles

for Organizations Developing Advanced AI systems

(draft for consultation)

The International Guiding Principles for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide and will provide guidance for organizations developing and using advanced AI systems, such as foundation models and generative AI. Organizations may include, among others, entities from academia, civil society, the private sector, and the public sector.

This non-exhaustive list of guiding principles will be discussed and elaborated as a living document to build on the existing OECD AI Principles in response to recent developments in advanced AI systems and are meant to help seize the benefits and address the risks and challenges brought by AI technologies. These principles should apply to all AI actors, when and as applicable to cover the design, development, deployment and use of advanced AI systems.

We look forward to developing these principles further as part of the comprehensive policy framework, with input from other nations and wider stakeholders in academia, business and civil society.

We also reiterate our commitment to elaborate an international code of conduct for organizations developing advanced AI systems based on the guiding principles below.

Different jurisdictions may take their own unique approaches to implementing these guiding principles 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.

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 that violate human rights, undermine democratic values, are particularly harmful to individuals or communities, facilitate terrorism, enable 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 principles, 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, mitigate risks across the AI lifecycle.

This includes employing diverse internal and independent external testing measures for advanced AI systems, through a mixture of methods 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 robustness, 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.

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, 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. Organizations are further encouraged to maintain appropriate documentation of reported incidents and to mitigate the identified risks and vulnerabilities, in collaboration with other

stakeholders. Reporting mechanisms to identify 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.

This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems.

Organizations should make the information in the transparency reports sufficiently clear and understandable to enable deployers and users to interpret the system's output and to enable users to use it appropriately, and that transparency reporting is supported and informed by robust internal documentation processes.

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

5 Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures, in particular for organizations developing advanced AI systems.

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 throughout the AI lifecycle.

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, operational security measures for information security and appropriate cyber/physical access controls.

7. Develop and deploy reliable content authentication and provenance mechanisms such as watermarking or other techniques to enable users to identify AI-generated content

This includes content authentication such as watermarking and/or provenance mechanisms for content created with an organization's advanced AI system. The watermark or provenance data should include an identifier of the service or model that created the content, but without including 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.

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 addressing key risks, as well as investing in developing appropriate mitigation tools.

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.

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

This includes contributing to the development and, where appropriate, use of international technical standards and best practices, including for watermarking, and working with international SDOs.

11 Implement appropriate data input controls and audits.

Organizations should commit to implementing appropriate safeguards throughout the AI lifecycle, and particularly before and throughout training, on the use of personal data, material protected by intellectual property rights including copyright-protected content, and other data which could result in harmful model capabilities. Appropriate transparency of training datasets should also be supported and organizations should comply with applicable legal frameworks.