

APPENDIX

Australia

Australia's approach to governance of generative AI has been led by senior figures in the Australian Government and reflects a measured and consultative process: commissioning expert reports, conducting public consultations, and coordinating across regulatory agencies.

These efforts aim to develop a risk-based governance framework proposing to permit low-risk generative AI applications while ensuring rigorous safeguards for high-risk use cases.

Coupled with guidance from bodies like the eSafety Commissioner, this balanced approach focuses on promoting innovation while mitigating potential harms through increased transparency, user protections, and industry responsibility.

AI Ethics Framework (November 2019)

Australia's **AI Ethics Framework**¹ was published in November 2019.

The AI Ethics Framework provides guidance to businesses and government entities on the responsible design, development, and implementation of AI. The Framework comprises 8 voluntary **AI Ethics Principles** that aim to ensure the safety, security, and reliability of AI applications and are intended to serve as best practices, complementing existing AI regulations and practices rather than replacing them.

Australia's AI Ethics Principles are entirely voluntary and are intended to encourage organizations to assess the implications of employing AI-enabled systems. The applicability of the AI Ethics Principles comes into play when the AI system, under development or implementation, is utilized to make decisions or significantly impacts people (including categorized groups), the environment, or society — whether positively or negatively. In cases where the developer is uncertain about how the AI system may impact its categorized groups or customers/clients, the AI Ethics Principles become applicable. However, it may not be necessary to consider all 8 of the principles if the AI use does not involve or affect human beings.

| AI Ethics Principle | Elaboration |
|---|---|
| Human, societal, and environmental wellbeing | AI systems should benefit individuals, society and the environment. |
| Human-centered values | AI systems should respect human rights, diversity, and the autonomy of individuals. |
| Fairness | AI systems should be inclusive and accessible and should not involve or result in unfair discrimination against individuals, communities or groups. |
| Privacy protection and security | AI systems should respect and uphold privacy rights and data protection and ensure the security of data. |
| Reliability and safety | AI systems should reliably operate in accordance with their intended purpose. |
| Transparency and explainability | There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI and can find out when an AI system is engaging with them. |
| Contestability | When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system. |
| Accountability | People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems and human oversight of AI systems should be enabled. |

Chief Scientist's Rapid Response Information Report on Generative AI (March 2023)

On 24 March 2023, Australia's Chief Scientist released a "Rapid Response Information Report" on Generative AI.² The Report was commissioned by Australia's National Science and Technology Council at the request of the Minister for Industry and Science, Ed Husic, in February 2023.

This Report has been cited in all subsequent policy documents released by the Australian Government on generative AI (see below).

The Report aims to answer the following questions:

- » What are the opportunities and risks of applying large language models (LLMs) and multimodal foundation models (MFMs) learning technologies over the next two, 5 and ten years?

- » What are some examples of strategies that have been put in place internationally by other advanced economies since the launch of models like ChatGPT to address the potential opportunities and impacts of artificial intelligence (AI)?

Based on a review of the existing literature, the Report provides a brief overview of how LLMs and MFMs function, the development landscape for these technologies, and highlights risks and opportunities for the Australian economy from use of these technologies.

In particular, the Report highlights the following **risks** from generative AI. The Report does not go so far as to propose regulatory measures to address these risks. However, it does highlight **potential solutions** that industry and/or regulators could implement to address certain of these risks.

| Risk | Potential Solution Highlighted |
|--|--|
| Factually inaccurate responses. | Ensuring that LLMs cite genuine sources and provide sufficient reasoning for their results. |
| Biased responses. | - |
| Spreading misinformation. | - |
| Lack of transparency for users and regulators as to how generative AI systems function. | Implementing a "human-in-the-loop" to ensure accountability and fairness, where appropriate. Conducting risk assessments, and developing mitigation strategies, including providing users with access to remedies. |
| Lack of transparency as to the datasets used to train generative AI models. | Obtaining consent for use of personal data in training datasets. Implementing privacy management programs for training datasets. Clarifying ownership of training datasets. Developing frameworks for sharing and using data, especially from public systems (e.g., in healthcare and education). |
| Data breaches, including through adversarial practices (e.g., 'jailbreaking'). | Security. |

Lastly, the Report also summarizes existing international strategies that aim to address these opportunities and risks and suggests future considerations.

Public Consultation on Safe and Responsible AI in Australia (June 2023 – January 2024)

On 1 June 2023, Australia's Department of Industry, Science and Resources (DISR) commenced a public consultation on how the Australian Government could

mitigate potential risks from AI and support safe and responsible AI practices.³ To guide the public consultation, the DISR released a discussion paper, titled "**Safe and Responsible AI in Australia**."⁴

Drawing on examples of regulatory efforts to govern AI internationally, the Discussion Paper sought input on potential governance and regulatory approaches to manage the risks of AI with the aim of increasing community trust and confidence in AI. This discussion also applies to AI generally, rather than generative AI in particular.

To that end, the Discussion Paper focuses mainly on presenting a spectrum of potential regulatory responses that the Australian Government could implement, ranging from releasing voluntary principles and guidelines to enacting or amending legislation, to address the risks of AI.

However, it does not specifically identify these risks or propose measures targeting specific risks. Rather, in Appendix C, the Discussion Paper presents a list of potential governance mechanisms that organizations could generally implement as part of a **risk-based approach** to the development and deployment of AI.

These mechanisms include:

- » **Impact assessments.**
- » **Notices** regarding how AI systems may materially affect users.
- » **Human-in-the-loop or oversight assessments.**
- » **Explanations** as to how AI systems arrive at specific outcomes or make decisions.
- » **Training** employees in the design, function and implementation of AI systems, so that employees can better identify and mitigate risks and explain and oversee operation of these systems.

- » **Monitoring and documentation** of an AI system, to ensure that they operate as intended and to identify and rectify any adverse or unintended impacts.

On 17 January 2024, the Australian Government published its “**Interim Response**” to the DISR’s consultation on safe and responsible AI in Australia.⁵

Broadly, the Interim Response reflects a risk-based approach to AI governance that aims to permit the use of AI in low-risk contexts while ensuring that the development and deployment of AI systems in legitimate but high-risk settings is safe and reliable.

Notably, the Interim Response acknowledges that many of the submissions received by the Australian Government focused on new risks posed by generative AI, including emerging ‘frontier models.’

Based on these submissions, the Interim Response identifies potential harms from AI systems and organizes them according to the three different stages of the AI product lifecycle: (1) development; (2) deployment; and (3) use.

| Stage of AI Product Lifecycle | Risk |
|---|--|
| Development, including the design and training of AI models. | Poor data governance resulting in inappropriate outputs. |
| | Use of inappropriate or biased data in model training. |
| | Data privacy. |
| | Ownership of data, including intellectual property. |
| Deployment, including release of AI models and integration of AI models into applications. | Competition issues. |
| Use, including outputs from AI models and actions by humans based on those outputs. | Consumer harms. |
| | Discrimination and bias. |
| | Lack of trust and transparency. |
| | Professional breaches. |
| | Misinformation and disinformation. |
| | Harmful content. |

The Interim Response also summarizes the submissions’ proposals for potential regulatory action to address these risks generally, without identifying solutions to specific risks. Potential regulatory actions cited in the Interim Report include strengthening existing laws and establishing ex-ante regulation, while non-regulatory actions include establishing an AI Advisory Body and regulatory sandboxes to support AI innovation, engaging in international AI governance initiatives, and building domestic AI capacity.

Notably, the Interim Response does not necessarily endorse or commit to implementing these actions.

Rather, the Interim Response indicates that in the short term, the Australian Government will continue to focus on implementing “soft law” mechanisms, such as a voluntary AI Safety Standard and establishing a temporary expert advisory group, and updating existing subject matter-specific regulation to address known harms of AI.

In the longer term, the Government is considering mandatory guardrails to address risks in legitimate but high-risk contexts. These potential guardrails focus on three areas: (1) testing; (2) transparency; and (3) accountability.

| Area | Potential Guardrail |
|----------------|---|
| Testing | Internal and external testing of AI systems before and after release (e.g., by independent experts). |
| | Sharing information on best practices for safety. |
| | Ongoing auditing and performance monitoring of AI systems. |
| | Cyber security and reporting of security-related vulnerabilities in AI systems. |
| Transparency | Letting users know when an AI system is used and/or that content is AI generated, including through labeling or watermarking. |
| | Public reporting on AI system limitations, capabilities, and areas of appropriate and inappropriate use. |
| | Public reporting on the data a model is trained on and sharing information on data processing and testing. |
| Accountability | Having designated roles with responsibility for AI safety. |
| | Requiring training for developers and deployers of AI products in certain settings. |

eSafety Commissioner's Tech Trends Position Statement on Generative AI (August 2023)

On 15 August 2023, Australia's eSafety Commissioner released⁶ its "**Tech Trends Position Statement on Generative AI**."⁷ The document is part of a broader series of statements of the eSafety Commissioner's position on emerging technologies, which include (among others) end-to-end encryption, recommender systems, and deep fakes,⁸ and draws on consultations with relevant stakeholders.

The Position Statement serves two main functions:

- » **explaining** how generative AI technologies function, and risks and opportunities from the technologies in the context of online safety; and
- » **providing guidance**, including the eSafety Commissioner's approach to governing technology, and recommendations for industry.

The Position Statement identifies the following risks to online safety from generative AI:

- » **Creation of abusive material**, including child sexual exploitation material, and material that radicalizes viewers or incites violence.

- » **Exposing minors to inappropriate content.**
- » **Encouraging or facilitating behavior that negatively impacts users' wellbeing and safety.**
- » **Creating non-consensual explicit material.**
- » **Facilitating cybercrime**, such as fraud.
- » **Facilitating harassment and bullying.**
- » **Generating content that reinforces stereotypes and amplifies existing biases.**
- » **Leaking personal data, including generating misleading or inaccurate information about individuals.**

The Position Statement also provides non-binding **recommendations** on "**good practices**" that industry could consider implementing to minimize the risk of harm throughout the lifecycle of a generative AI system's recommendations. These recommendations are based on three "Safety by Design" principles: (1) Service Provider Responsibility; (2) User Empowerment and Autonomy; and (3) Transparency and Accountability.

| Safety by Design Principle | Practice |
|---------------------------------|---|
| Service Provider Responsibility | Making teams accountable for safety , including creating, implementing, operating, and evaluating user safety policies, and promoting a culture of safety as a whole. |
| | Having policies and procedures to prevent harms before they occur , including: <ul style="list-style-type: none"> » Risk and impact assessments to assess and remediate harms. » Prompt testing and design, including automated and manual tests and creative testing of edge cases. » Red teaming and violet teaming. » Data collection and curation, including consideration of privacy obligations, and data ethics, consent, ownership, and provenance. » Ongoing evaluation and continuous improvement of systems. |
| | Age-appropriate design , supported by robust age assurance measures, to identify minors and apply age-appropriate safety and privacy settings. |
| | Internal protocols for working with law enforcement, support services and illegal content hotlines. |
| | Digital watermarking of AI-generated content. |
| | Establishing a system to handle user safety concerns , including making it easy for people to report concerns and violations as soon as they happen. |
| | |
| User Empowerment and Autonomy | Clearly outlining the rights, responsibilities, and safety expectations for the service, users, and third parties. |
| | Using technical interventions to educate and empower users , including: <ul style="list-style-type: none"> » Implementing informed consent for collection and use of users' data. » Providing disclaimers and content warnings to let users know that outputs could be incorrect, biased, or harmful. » Developing educational content about how to detect AI 'hallucinations' or other forms of false or harmful content. » Providing users opportunities to understand, evaluate, control, and moderate their own interactions (e.g., real-time prompts and nudges to alert users to safety features). |
| | Providing real-time support and enabling user reporting. |
| Transparency and Accountability | Providing clear and accessible information about user safety policies, privacy policies, terms and conditions, community guidelines, and processes. |
| | Innovating and investing in new technologies to enhance user safety. |
| | Consulting with a diverse user base through open engagement and engaging with experts who have specialist knowledge in various forms of harm. |
| | Publishing regular transparency reports about reported abuses and meaningful analysis of metrics. |
| | Documenting the capabilities, limitations, intended uses and prohibitive uses of AI models (for example, through model cards, system cards, and value alignment cards). |
| | Considering granting independent researchers, academics access to models. |

Digital Platform Regulators Forum Working Paper on LLMs (October 2023)

Unlike the eSafety Commissioner in relation to online safety, Australia's federal data protection authority, the Office of the Australian Information Commissioner (OAIC), has not released any guidance on the application of Australia's data protection law, the Privacy Act 1988, to generative AI systems.

However, on 23 October 2023, the Digital Platform Regulators Forum (DP-REG)⁹ – which includes the OAIC as well as Australian Competition and Consumer Commission, the Australian Communications and Media Authority, and the eSafety Commissioner – released¹⁰ a working paper examining LLMs and their impact on the regulatory roles of each member of the DP-REG.¹¹

The working paper identifies the following risks that may arise from the deployment of generative AI.

| Regulatory Domain | Risks |
|---------------------------------------|--|
| Consumer Protection | Facilitating scams, fake reviews and harmful applications , by creating more convincing forms of fraudulent content at scale and enabling threat actors without sophisticated programming skills to create malware. |
| | Creating misleading and deceptive content. |
| Competition | Making it harder for new entrants to compete with digital platform services that use LLMs , as large digital platforms may have advantages in data, computing power, financial resources, economies of scale and 'positive feedback loops.' |
| | Potentially increasing anti-competitive conduct , such as self-preferencing, typing, and data access restriction. |
| Media and the Information Environment | Reinforcing and reproducing biases present in their training data. |
| | Facilitating the spread of misinformation , whether accidentally or through malicious use. |
| | Producing inaccurate or out-of-date information. |
| Privacy | Lack of transparency in the processing of personal data. |
| | Disclosure of inaccurate personal data. |
| | Lack of control for data subjects over use of their personal data , especially where training datasets have been scraped from public websites without data subjects' knowledge or consent. |
| | Data breach. |
| | Creation of personalized content for manipulative purposes. |
| Online Safety | Abuse, bullying, harassment and hate at scale. |
| | Manipulation, impersonation, and exploitation. |
| | Age-inappropriate content. |

China

China's approach to the governance of generative AI aims to cultivate a generative AI ecosystem aligned with state interests and socialist principles.

China's comprehensive and multi-layered regulatory approach to generative AI reflects the government's firm stance on harnessing these powerful technologies to drive economic and technological development, while enforcing strict oversight and content controls to eliminate perceived threats to national security and public order. In particular, enhanced obligations for services capable of swaying public discourse demonstrate the paramount priority placed on controlling narratives and information flows.

However, such restrictive governance could also stifle research and commercialization if implemented overzealously. Striking this balance will likely remain an ongoing challenge for Chinese authorities as generative AI capabilities rapidly evolve.

Ethical Principles for New Generation AI (September 2021)

On 25 September 2021, the National New Generation AI Governance Specialist Committee within China's Ministry of Science and Technology (MOST) released a set of "**Ethical Principles for New Generation AI**" (新一代人工智能伦理规范)(Ethical Principles).¹²

These Principles are intended to provide guidance to persons and organizations on incorporating ethics into the entire lifecycle of an AI system. They implement the:

- » "**New Generation Artificial Intelligence Development Action Plan**" – a top-level design blueprint released by China's State Council in 2017 outlining China's national approach to the development and application of AI technology, as well as broad goals up to 2030.¹³
- » "**Governance Principles for a New Generation of Artificial Intelligence**" – a set of eight high-level principles for AI governance and responsible AI released by the MOST's National New Generation AI Governance Specialist Committee in 2019.¹⁴

The Ethical Principles establish six basic ethical principles that apply to all AI-related activities:

| | |
|--|--|
| Advancement of human welfare (增进人类福祉) | <p>AI-related activities should be human-centric and abide by shared human values, respect human rights and appeals to fundamental human interests, and comply with national or regional ethics.</p> <p>AI-related activities should prioritize the public interest; promote human harmony and friendship; improve the people's livelihoods, improve people's livelihoods and happiness; advance sustainable economic, social, and ecological development, and jointly build a community of common destiny for humanity.</p> |
| Promotion of fairness and justice (促进公平公正) | <p>AI-related activities should uphold inclusivity and tolerance; safeguard the legitimate rights and interests of each relevant entity; promote fair sharing of AI benefits throughout society; and promote social equity, justice, and equal opportunities.</p> <p>When providing AI products and services, AI actors should fully respect and help vulnerable groups and special groups, and provide appropriate alternatives as necessary.</p> |

| Principle | Elaboration |
|---|---|
| Protection of privacy and security (保护隐私安全) | <p>AI-related activities should fully respect everyone's right to know the extent of the use of, and to consent to the use of, their personal data.</p> <p>AI actors should process personal data according to the principles of legality, propriety, necessity, and good faith, and guarantee personal privacy and data security.</p> <p>AI actors should not harm individuals' legal data rights and interests; steal, tamper, leak, or otherwise illegally collect or use personal data; or infringe upon personal privacy rights.</p> |
| Assurance of controllability and trustworthiness (确保可控可信) | <p>AI actors should ensure that humans are granted the rights to make fully autonomous decisions; accept or reject AI-provided services; withdraw from AI interactions at any time; and terminate AI system operations at any time.</p> <p>AI actors should also ensure that AI is always under human control.</p> |
| Strengthening accountability (强化责任担当) | <p>AI actors should clearly define the responsibilities of relevant parties; increase parties' awareness of these responsibilities and exercise self-reflection and self-discipline at every stage of the AI life cycle.</p> <p>AI actors should also establish AI accountability mechanisms and should not avoid investigations into responsibility or evade their own responsibilities.</p> |
| Improving ethical literacy (提升伦理素养) | <p>AI actors should actively learn about and spread awareness of AI ethics.</p> <p>AI actors should objectively understand ethical issues and should not underestimate or exaggerate ethical risks.</p> <p>AI actors should actively carry out or participate in discussions of AI ethical issues.</p> <p>AI actors should thoroughly promote the practice of AI ethical governance and improve their ability to respond to ethical issues.</p> |

In addition to outlining broad ethical principles that apply to all AI-related activities, China's Ethical Principles also provide more granular requirements for specific activities, including:

- » **Management**, which is defined to include AI-related strategic planning, developing and implementing policies, regulations, and technical standards; allocating resources; and supervision and examination.
- » **Research and development**, which are defined to include scientific research, technological development, and product development relating to AI.
- » **Supply activities**, which are defined to include AI product and service-related production, operations, and sales.
- » **Use activities**, which are defined to include purchasing, consuming, and operating AI related products and services.

| Activity | Responsibility | Elaboration |
|------------|---|--|
| Management | Promoting agile governance | <p>Persons involved in the management of AI-related activities should respect the laws governing the development of AI, fully understand the potential and limitations of AI, and continuously optimize governance mechanisms and approaches.</p> <p>In the processes of strategic decision-making, establishing institutions, and allocating resources, persons involved in the management of AI-related activities should promote healthy, sustainable, and orderly development of AI without departing from reality or seeking short-term gains.</p> |
| | Actively practicing ethics and demonstrating how to put ethics into practice | <p>Persons involved in the management of AI-related activities should comply with relevant laws, policies, and standards relating to AI and actively integrate ethical considerations into the entire management process.</p> <p>They should become pioneers and promoters of ethical AI governance, promptly disseminate summaries of their experiences with AI governance, and actively respond to societal concerns regarding AI ethics.</p> |
| | Correctly exercising authority | <p>Persons involved in the management of AI-related activities should define the responsibilities for AI related management activities and identify the limits of each parties' authority.</p> <p>They should establish conditions and procedures for the exercise of authority and fully respect and safeguard the privacy, freedom, dignity, and security rights, and other legitimate rights and interests of relevant entities.</p> <p>They should also prohibit improper exercises of authority that may harm the legitimate rights and interests of natural persons, legal persons, and other organizations.</p> |
| | Strengthening risk prevention | <p>Persons involved in the management of AI-related activities should improve their baseline thinking and awareness of risks, assess potential risks in the development of AI, and conduct timely and systematic risk monitoring and evaluation.</p> <p>They should also establish effective warning mechanisms and improve their capabilities to control and handle ethical risks.</p> |
| | Promoting inclusive openness | <p>Persons involved in the management of AI-related activities should give full consideration to the rights and expectations of all stakeholders in AI.</p> <p>They should encourage the application of diversified AI technologies to address practical economic and social development issues.</p> <p>They should also promote interdisciplinary, cross-domain, cross-regional, and international exchanges and cooperation, facilitating the formation of widely accepted frameworks and standards for AI governance.</p> |

| Activity | Responsibility | Elaboration |
|--------------------------|--|--|
| Research and development | Strengthening “self-discipline” | <p>Persons involved in researching and developing AI should exercise self-restraint in AI-related research and development.</p> <p>They should actively incorporate ethical considerations into each stage of the research and development process, conduct self-examination conscientiously, strengthen self-management, and refrain from engaging in unethical AI research and development.</p> |
| | Improving data quality | <p>Persons involved in researching and developing AI should strictly comply with data-related laws, standards, and regulations when collecting, storing, using, processing, transmitting, providing, and disclosing data.</p> <p>They should also improve the integrity, timeliness, consistency, standardization, and accuracy of data.</p> |
| | Enhancing security and transparency | <p>Across the algorithm design, implementation, and application stages, persons involved in researching and developing AI should:</p> <ul style="list-style-type: none"> » strengthen AI systems’ capabilities for resilience, adaptability, and anti-interference; and » enhance the transparency, interpretability, understandability, reliability, and controllability of AI systems; » gradually achieving verifiability, auditability, supervisability, traceability, predictability, and reliability of AI systems. |
| | Avoiding bias and discrimination | <p>When collecting data and developing algorithms, persons involved in researching and developing AI should strengthen ethical review and consider the different needs of various kinds of users.</p> <p>They should also avoid potential data and algorithm biases, and strive to achieve inclusiveness, fairness, and non-discrimination in AI systems.</p> |

| Activity | Responsibility | Elaboration |
|----------|---|---|
| Supply | Respecting market rules | <p>Persons involved in supplying AI-related products and services should strictly comply with regulations governing market access, competition, and transactions.</p> <p>They should actively maintain market order, create a market environment conducive to the development of AI, and prohibit the disruption of market order through data or platform monopolies. Prohibit any means that infringe on the intellectual property rights of other entities.</p> |
| | Strengthening quality control | <p>Persons involved in supplying AI-related products and services should strengthen the quality monitoring and usage assessment of AI products and services, avoiding harms to health, property, user privacy, and similar interests that may be caused by design and product defects.</p> <p>They should not operate, sell, or provide products and services that do not meet quality standards.</p> |
| | Safeguarding users' rights and interests | <p>Persons involved in supplying AI-related products and services should clearly inform users, identifying the functions and limitations of the products and services.</p> <p>They should guarantee that users have the right to be informed about and consent to the use of products and services and should provide simple and understandable solutions for users to choose to use or opt-out of AI modes.</p> <p>They also should not create barriers to the equal use of AI by users.</p> |
| | Strengthening emergency support | <p>Persons involved in supplying AI-related products and services should research and formulate emergency mechanisms and plans or measures to compensate users for losses.</p> <p>They should monitor AI systems in a timely manner, respond promptly to and handle user feedback, prevent systemic failures in a timely manner, and be ready to assist relevant entities in intervening in AI systems in accordance with the law and regulations, reducing losses and avoiding risks.</p> |

| Activity | Responsibility | Elaboration |
|----------|--|--|
| Use | Promoting ethical use of AI | Users of AI should strengthen pre-use demonstrations and assessments of AI products and services. They should gain a full understanding of the benefits of AI products and services, and give full consideration to the legitimate rights and interests of all stakeholders. They should also effectively promote economic prosperity, social progress, and sustainable development. |
| | Avoiding misuse and abuse | Users of AI should fully understand the scope of applications and the negative impacts of AI products and services. They should respect the right of relevant entities not to use AI. They should also avoid improper use and abuse of AI products and services and prevent unintentional harm to the legitimate rights and interests of others. |
| | Prohibiting illegal and malicious use | Users of AI should prohibit the use of AI products and services that do not comply with laws, regulations, ethics, standards, and norms. They should also prohibit the use of AI products and services for illegal activities and strictly prohibit actions that endanger national security, public safety, and production safety, or harm public interests. |
| | Providing timely and proactive feedback | Users of AI should actively participate in the practice of ethical AI governance. They should also provide timely feedback to relevant entities on discovery of technical security vulnerabilities, policy and regulatory vacuums, and lagging supervision during the use of artificial intelligence products and services, and should assist in solving these issues. |
| | Improving abilities to use AI | Users of AI should actively develop AI-related knowledge, proactively master the skills required for operating and maintaining AI products and services and responding to emergencies, ensuring the safe and efficient use of AI products and services. |

Regulations on the Administration of Deep Synthesis of Internet Information Technology (January 2023)

The Cyberspace Administration of China (CAC)'s **"Regulations on the Administration of Deep Synthesis of Internet Information Technology"** (互联网信息服务深度合成管理规定)¹⁵ (Deep Synthesis Regulations) were enacted on 3 November 2022 and entered into force on 10 January 2023.

Scope and Enforcement

These Regulations apply to the online provision of services that use **"deep synthesis technology"** in the People's Republic of China (deep synthesis services).

"Deep synthesis technology" refers to technologies that use deep learning, virtual reality and other forms of generative sequencing algorithms to generate and

edit various forms of content, including text, voice, music and sound, images, biometric data (e.g., face, posture), digital characters and virtual scenes.

The Regulations are legally binding rather than voluntary.

They impose obligations on organizations and individuals that:

- » provide deep synthesis services (Service Providers);
- » provide technical support for deep synthesis services (Technical Supporters);
- » use deep synthesis services to make, reproduce, publish, or transmit information (Users),

as well as application distribution platforms (App Platforms).

They also empower relevant authorities to conduct supervision and inspections of deep synthesis services and impose penalties on Service Providers and Technical Supporters under relevant laws and regulations.

Prohibition on Certain Uses of Deep Synthesis Technology

The Regulations expressly prohibited use of deep synthesis services for:

- » producing, reproducing, publishing, or transmitting illegal information, or engaging in illegal activities.

Under Chinese law, activities or content may be considered illegal if they:

- » endanger national security and interests;
- » harm the image of the nation;
- » harm societal public interest;
- » disturb economic or social order; or
- » harm the lawful rights and interests of others.

The Deep Synthesis Regulations prohibit use of technical means to delete, tamper with, or conceal watermarking as required by the Regulations.

Obligations

The majority of obligations under the Deep Synthesis Regulations apply to Service Providers. Service Providers that are in a position to alter public opinion or mobilize the public are also required to register with relevant regulators. They must also conduct a security assessment before launching new products, applications, or features that may alter public opinion or mobilize the public.

| Actor | Obligations |
|-------------------|---|
| Service Providers | Undertaking primary responsibility for information security , and reminding Technical Supporters and Users of their information security obligations. |
| | Establishing and improving management systems for: <ul style="list-style-type: none"> » user registration, » scientific and technological ethical review, » information release review, » data security, » protection of personal data, » combatting telecommunication network fraud, » emergency response, and other management systems. |
| | Implementing safe and controllable technical safeguards . |
| | Publishing management rules and platform conventions . |
| | Implementing user verification , and prohibiting access to users who do not provide genuine identity information. |
| | Establishing and strengthening content management and review measures. |
| | Reporting illegal or undesirable content to relevant authorities, and sanctioning relevant Users according to law. |
| | Establishing mechanisms to identify and debunk misinformation , and reporting the misinformation to relevant authorities. |
| | Establishing a system for user appeals, public complaints, and reports . |
| | Strengthening the management and security of training data . |
| | Notifying and obtaining consent from data subjects , where the service enables editing of their biometric information (e.g., faces and voices). |
| | Conducting security assessments where services enable generation or editing of biometric information or content that might involve national security, the nation's image, national interests, and the societal public interest. |
| | Watermarking content produced and edited by Users. |
| | Prominently labeling content as potentially misleading if services involve: <ul style="list-style-type: none"> » Simulated text generation or editing, such as intelligent conversations and intelligent writing. » Voice synthesis, mimicry, or significant alteration of personal identity features in voice editing services. » Face generation, face replacement, face manipulation, posture manipulation, and other image or video editing services significantly altering personal identity features. » Immersive and realistic scene generation or editing services. » Other services with functions significantly altering information content. |

| Actor | Obligations |
|----------------------|--|
| Technical Supporters | Strengthening the management and security of training data. |
| | Notifying and obtaining consent from data subjects , where the service enables editing of their biometric information (e.g., faces and voices). |
| | Conducting security assessments where services enable generation or editing of biometric information or content that might involve national security, the nation's image, national interests, and the societal public interest. |
| App Platforms | Implementing safety mechanisms , such as pre-offering reviews, routine management, and emergency response. |
| | Checking deep synthesis services' security assessments and filings . |
| | Promptly employing measures to address any violation of state provisions . |

Interim Measures for the Management of Generative AI Services (August 2023)

The Cyberspace Administration of China (CAC)'s **"Interim Measures for the Management of Generative AI Services"**(生成式人工智能服务管理暂行办法) (Interim Generative AI Measures)¹⁶ were enacted on 10 July 2023 and took effect on 15 August 2023.

The Interim Generative AI Measures are more extensive and detailed than the Deep Synthesis Regulations and cover broader subject matter. Rather than simply assigning administrative responsibility for the supervision of generative AI, they contain a broader statement of state policy in relation to generative AI, highlighting the opportunities presented by generative AI, and outlining generative AI-specific principles.

Broadly, these principles require providers and users of generative AI services to:

- » adhere to state values, and refrain from creating content that undermines the state, or that promotes ethical discrimination, violence, obscenity, or the spread of false or harmful information;
- » take effective measures to prevent discrimination on the basis of factors such as race, religion, country, region, gender, age, occupation, health, in the design of algorithms, the selection of training data, the creation and optimization of models, and the provision of services;

- » respect intellectual property rights and business ethics, keep trade secrets, and refrain from monopolization and unfair competition using algorithms, data, platforms, and other advantages;
- » respect the legitimate rights and interests of others, and avoid infringing on the rights to image, reputation, honor, privacy, and personal information of others.
- » based on the nature of the service, take effective measures to enhance the transparency of generative AI services and improve the accuracy and reliability of generated content.

The Measures define **"generative AI"** as models and related technology that have the ability to generate content, such as text, images, audio, and video.

The Measures impose a variety of obligations on **"Generative AI Service Providers"** – defined as organizations or individuals that provide services using generative AI within the People's Republic of China – throughout the lifecycle of a generative AI system.

As with the Deep synthesis Regulations, Generative AI Service Providers that are in a position to alter public opinion or mobilize the public are subject to stricter obligations. These include:

- » conducting security assessments according to relevant national regulations, and
- » fulfilling the requirements of algorithm filing, changes and cancellation filing procedures according to the Measures for the Management of Internet Information Service Algorithm Recommendation.

| Stage | Obligation on Service Providers |
|-------------------------------|---|
| Training generative AI models | Using data and basic models from legal sources. |
| | Refraining from infringing upon others' legal rights over their intellectual property. |
| | Obtain consent for use of others' personal data or otherwise satisfying another legal basis for processing such data. |
| | Taking effective measures to improve the quality, and enhance the authenticity, accuracy, objectivity and diversity of the training data. |
| | Comply with relevant laws and regulations. |

| Stage | Obligation on Service Providers |
|---|--|
| Annotating data in the development of generative AI systems | Formulating clear, specific, and operational annotation rules. |
| | Conduct quality assessments of data annotation. |
| | Conducting randomized checks on the accuracy of annotated content. |
| | Providing necessary training to annotation personnel to enhance their awareness of obligations under relevant laws and regulations. |
| | Supervising and guiding annotation personnel in carrying out annotation work in a standardized manner. |
| After deployment | Entering into service agreements with users to clarify the rights and obligations of both parties. |
| | Clearly and publicly state the applicable target audience, occasions, and purposes of their services. |
| | Guiding users to understand and use generative AI technology rationally. |
| | Taking effective measures to prevent minors from excessively relying on or becoming addicted to generative AI services; |
| | Regarding personal data , <ul style="list-style-type: none"> » protect information inputted by users, and users' usage records according to relevant laws, such as the Personal Information Protection Law; » avoid collecting unnecessary personal data, unlawfully retaining input information and usage records that can identify the user's identity, or unlawfully providing such information to others; » promptly handle and process requests from individuals regarding the inquiry, copying, correction, supplementation, or deletion of their personal information in accordance with the law. |
| | Watermarking generated content. |
| | Ensuring the security, stability, and continuity of their services during the service process, ensuring users' normal usage. |
| | On discovering illegal content, promptly taking appropriate measures , such as: <ul style="list-style-type: none"> » stopping generation, transmission, and elimination; » implementing model optimization and training to rectify the situation; » reporting the content to relevant competent authorities. |
| | On discovering that users are engaging in illegal activities using generative AI services, taking appropriate measures , such as: <ul style="list-style-type: none"> » Issuing warnings, » Imposing functional restrictions, suspensions, or » Terminating services in accordance with the law and the service agreement; » Maintaining relevant records, and » Reporting the conduct to relevant competent authorities; |
| | Establishing sound complaint and reporting mechanisms, provide convenient channels for complaints and reports, publicize the handling process and feedback time limits, promptly accept and handle public complaints and reports, and provide feedback on the handling results. |

Enforcement and Remedies

Users who find that service providers have failed to comply with the Measures may lodge a complaint with relevant authorities.

The Measures empower relevant authorities to conduct supervision and inspection of generative AI services, implement technical measures to prevent overseas providers who do not comply with the Measures from providing services in the PRC, and subject service providers to penalties under relevant laws or regulations.

In the absence of penalties under other laws/regulations, the CAC may:

- » issue warnings;
- » circulate criticisms;
- » order corrections within a set period of time; or
- » where corrections are refused or circumstances are grave, order suspension of provision of generative AI provider services.

TC260's Basic Security Requirements for Generative Artificial Intelligence Services (February 2024)

On 29 February 2024, China's National Cybersecurity Standardization Technical Committee, also known as

TC260, released the “**Basic Security Requirements for Generative AI Services**” (生成式人工智能服务安全基本要求)¹⁷ – a technical standard that sets out the basic security requirements that service providers must follow under the Interim Generative AI Measures.

These Requirements are non-exhaustive – service providers are also expected to comply with other network and data security and data protection laws.

It also outlines criteria for detailed security assessments. These include testing training data against a database of at least 10,000 keywords and generated content against a bank of at least 2,000 test questions to detect the presence of 31 security risk types in 5 areas:

- » Content that violates the core values of socialism;
- » Discriminatory content;
- » Content that violates commercial laws and regulations (including intellectual property);
- » Infringing on the legitimate rights and interests of others; and
- » Inaccurate or unreliable content when services are provided in high security areas, such as medicine, psychological counseling, and critical information infrastructure.

| Stage | Obligation on Service Providers |
|-------------------------|---|
| Compiling training data | Illegal and negative information: Before collecting from specific corpus sources, a security assessment of the corpus should be conducted. The corpus should not be used if it contains more than 5% “illegal or harmful information” as defined in the “Regulations on Ecological Governance of Internet Information Content.” ¹⁸ Information blocked under China's cybersecurity laws, regulations, and policy documents should not be used to train generative AI models. Service providers should filter out illegal and harmful content from the training corpus using keywords, classification models, manual sampling and other methods. |
| | Diversification in sources of training data: Multiple sources of data should be used. If foreign corpora are used, they should be combined with domestic corpora. |
| | Traceability: Training data should be traceable. There should be a collection record, open-source license, or a legally enforceable contract for use of the data that contains commitments and relevant supporting materials as to the source, quality, and safety of the corpus. |
| | Intellectual property: Responsible personnel for corpus and generated content intellectual property rights should be designated, and an intellectual property rights management strategy should be established. Before training, major intellectual property infringement risks in the corpora should be identified. If there are issues such as intellectual property rights infringement, service providers should not use the related corpora for training. |
| | Use of personal data: If the corpus contains personal data, the service provider should obtain the data subject's consent for use of their personal data to train a generative AI model, unless another legal basis applies. |

| Stage | Obligation on Service Providers |
|--|--|
| Use of prompt data | <p>Data from user prompts should only be used to train a model if users have authorized such use.</p> <p>Convenient methods should be provided for users to opt-out of using their input information for training. The opt-out process should be straightforward and should involve no more than 4 clicks from the main interface.</p> |
| Annotating training data | <p>The Basic Security Requirements outline detailed requirements for training and qualification of personnel responsible for annotating training data, as well as conducting the annotation process. Different requirements apply for function annotation and security annotation.</p> |
| Using models developed by third parties | <p>Service providers who use models developed by third parties should use models that have been filed with the competent authority.</p> |
| Safety of generated content | <p>During the training process, the safety of generated content should be considered one of the main criteria for evaluating the quality of the generated results.</p> <p>In each conversation, the input information from users should undergo safety checks to guide the model to generate positive and constructive content.</p> <p>Monitoring and evaluation methods should be established to promptly address safety issues and optimize the model through targeted instruction fine-tuning, reinforcement learning and other methods.</p> <p>Technical measures should also be adopted to improve the accuracy and reliability of generated content.</p> |
| Use of generative AI in certain sectors | <p>It is necessary to fully demonstrate the necessity, applicability, and safety of using generative AI to provide services in various fields. Appropriate protections corresponding to the level of risk should be put in place when using generative AI to provide services used in critical information infrastructure, as well as important scenarios such as automatic control, medical information services, psychological counseling, financial information service.</p> |
| Minors | <p>For services applicable to minors:</p> <ul style="list-style-type: none"> » Guardians should be allowed to set anti-addiction measures for minors. » Paid services should not be provided to minors if the services are inconsistent with the legal capacity of minors. » Services should actively present content that is positive and beneficial for the physical and mental health of minors. <p>Technical or managerial measures should be taken to prevent minors from using services not applicable to minors.</p> |
| Transparency | <p>For services provided through interactive interfaces, information should be provided about:</p> <ul style="list-style-type: none"> » the applicable users, scenarios, purposes; » The limitations of the service; and » A summary of the generative AI model or algorithm used. <p>Information on whether user inputs are used to train the model, and how to opt out of this, should be prominently displayed.</p> |
| Supply chains | <p>The supply chain security of chips, software, tools, computing power, etc., adopted by the system should be evaluated, with a focus on assessing aspects such as supply continuity and stability.</p> <p>The adopted chips should support hardware-based secure boot, trusted boot processes, and security verification to ensure that generative artificial intelligence systems operate in a secure and trustworthy environment.</p> |

| Stage | Obligation on Service Providers |
|---|--|
| Complaints channels | Channels and feedback methods for receiving public or user complaints should be provided. Rules and deadlines for processing public or user complaints should be established. |
| Provision of services to users | <p>Detection of user input information should be conducted using methods such as keywords, classification models, etc. If a user inputs illegal and harmful information three times in a row or accumulates 5 times within a day, or induces the generation of such information, measures such as suspending service provision should be taken in accordance with the law and contracts;</p> <p>Questions that are evidently biased or induce the generation of illegal and harmful information should be refused to be answered; other questions should be answered normally.</p> <p>Monitoring personnel should be appointed, and the quality and security of generated content should be improved promptly based on monitoring. The number of monitoring personnel should be matched with the scale of the service.</p> |
| Model updates and upgrades | <p>Security management strategies should be formulated for model updates and upgrades.</p> <p>A management mechanism should be established to organize security assessments again after significant model updates or upgrades.</p> |
| Service stability and continuity | <p>The training environment should be isolated from the inference environment to prevent data leakage and unauthorized access.</p> <p>Continuous monitoring of model input content should be conducted to prevent malicious input attacks, such as DDoS, XSS, injection attacks, etc.</p> <p>Regular security audits should be conducted on the development frameworks, codes, etc., used, focusing on security issues and vulnerabilities related to open-source frameworks, identifying and fixing potential security vulnerabilities.</p> <p>Backup mechanisms and recovery strategies for data, models, frameworks, tools, etc., should be established, with a focus on ensuring business continuity.</p> |

Draft AI Law (March 2024)

On 31 May 2023, China's State Council released its legislative work plan for 2023.¹⁹ The plan briefly states that the Standing Committee of the National People's Congress has been requested to deliberate on a draft Artificial Intelligence Law, among various other items of draft legislation.

For context, the National People's Congress functions as China's national legislature. Its Standing Committee is a permanent body that exercises the

powers of the National People's Congress when it is not in session.

On 16 March 2024, an expert group comprising academics from several Chinese universities released an academic draft of the Artificial Intelligence Law at a symposium on "AI Good Governance Forum and Prospect of Artificial Intelligence Legal Governance" in Beijing.²⁰ It remains unclear whether the Chinese government will adopt this academic draft as national AI law in its current form or otherwise.

Japan

Japan's approach to governance of generative AI is based on voluntary cross-sector guidelines for ethical AI practice, and Japan has prioritized international cooperation to develop unified governance norms.

As G7 president in 2023, Japan has led international efforts to establish international standards around advanced AI systems, including generative models. Notably, Japan launched the **Hiroshima AI Process**, which aims to foster inclusive global governance for advanced AI. In December 2023, it Process produced

the first major international framework for advanced AI systems, comprising International Guiding Principles for all AI actors across the lifecycle, and an International Code of Conduct for organizations developing advanced AI systems.

In December 2023, Japan released for public consultation a set of draft AI governance guidelines that aim to update its AI governance framework to address generative AI and reflect progress made during the Hiroshima AI process.

Separately, Japan's data protection authority has engaged directly with privacy challenges from generative AI by issuing guidance in June 2023 on use of LLM chatbots under Japan's data protection law and pursuing enforcement against OpenAI regarding ChatGPT's handling of sensitive personal data.

Social Principles of Human-Centric AI (March 2019)

The Cabinet Office of Japan released the “**Social Principles of Human-Centric AI**” (人間中心の AI 社会原則)²¹ on 29 March 2019.

The Principles highlight the benefits of AI and call for transformation of the whole of Japanese society – including human resources, social systems, industrial structures, innovation, and governance – into an “**AI Ready Society**” that uses AI effectively while avoiding or reducing any negative aspects.

The Principles are based on three **basic values** that constitute an AI Ready Society:

| Basic Value | Elaboration |
|--|---|
| Dignity (人間の尊厳が尊重される社会) | A society that has respect for human dignity, where humans are not overly dependent on AI and AI is not used to control people but rather, where AI is a tool for people to demonstrate human abilities and creativity, engage in challenging works, and live richer lives physically and mentally. |
| Diversity and Inclusion (多様な背景を持つ人々が多様な幸せを追求できる社会) | A society where people with diverse backgrounds, values, and ways of thinking can pursue their own well-being while society creates new value by embracing them. |
| Sustainability (持続性ある社会) | A society that uses AI to create new businesses and solutions, resolve social disparities, and develop a sustainable society that can deal with issues such as global environmental problems and climate change. |

The Principles outline seven “Social Principles of AI” for all stakeholders in society to keep in mind to realize an AI-Ready Society:

| | |
|---|---|
| Human Centricity (人間中心) | In implementing AI, stakeholders should adhere to human rights and international standards, ensuring that AI enhances individual capabilities. The responsible development of AI involves literacy education to prevent over-dependence and misuse. AI's role is to augment human abilities and creativity, serving as an advanced tool rather than a replacement. Users must make informed decisions on AI usage, and stakeholders bear responsibility for consequences. AI deployments should prioritize user-friendliness, preventing a digital divide and ensuring equitable access to AI benefits for all, including those deemed “information poor” or “technology poor.” |
| Education/Literacy (教育・リテラシー) | In an AI-centric society, preventing social disparities is paramount. Policymakers and business managers in the AI field must accurately understand AI and AI ethics to ensure responsible use of AI in society and must appreciate the complexity of AI and its potential for misuse. Users of AI should also have a sufficient education in AI to use the technology appropriately. Developers should focus not only on technical skills but also business models for societal use of AI and social sciences and ethics. The educational environment for AI must be equitable and principled, creating opportunities for people of all ages and across multiple domains. |

| Social Principles of AI | Elaboration |
|--|--|
| Privacy Protection (プライバシー確保) | Because AI technologies can accurately assess individuals' characteristics based on their behavior, personal data must be carefully handled to prevent harm in an AI society. Stakeholders must avoid infringing personal freedom, dignity, and equality. Technical and non-technical measures should mitigate risks associated with AI use, particularly in handling personal data. AI systems should prioritize accuracy, legitimacy, and individual involvement in privacy management. Protection of personal data must align with its importance and sensitivity, considering a broad range of information. Striking a balance between data use and protection is essential, respecting cultural backgrounds and societal norms. |
| Ensuring Security (セキュリティ確保) | Active AI use automates many social systems and improves safety but introduces security risks, as AI may not adequately respond to rare events or intentional attacks. Societal awareness of the balance between AI benefits and risks is crucial, emphasizing continuous efforts to improve overall safety and sustainability. To address this, broad and in-depth research on AI, including risk assessment and mitigation strategies, is essential. Risk management, especially in cybersecurity, should be a priority. Additionally, society should avoid over-reliance on specific AI types to ensure sustainability in AI utilization. Ongoing vigilance and comprehensive measures are necessary for responsible AI integration into society. |
| Fair Competition (公正競争確保) | Maintaining a fair competitive environment is crucial for fostering new businesses, maintaining sustainable economic growth, and addressing societal challenges. Regardless of the concentration of AI resources in a country or specific companies, it is essential to prevent unfair data collection, infringement of sovereignty, and biased wealth distribution. Societal frameworks should discourage dominant positions leading to unjust competition and ensure that the use of AI promotes equitable wealth distribution and social influence among stakeholders. This approach safeguards against imbalances, fostering a fair and inclusive landscape for the development and deployment of AI technologies. |
| Fairness, Accountability, and Transparency (公平性、説明責任及び透明性) | An "AI-Ready Society" demands fairness, transparency, and accountability in decision-making and should aim to prevent discrimination based on personal background and uphold human dignity. The design concept of AI should treat everyone fairly, irrespective of factors like race, gender, nationality, age, or beliefs. Detailed explanations about AI applications, data usage, and result appropriateness must be provided case by case. Open dialogues are crucial to enable public understanding and judgment of AI proposals. To safely integrate AI into society, a trustworthy mechanism encompassing both AI and its supporting data and algorithms should be established, ensuring confidence in the technology and fostering societal acceptance. |
| Innovation (イノベーション) | Achieving Society 5.0 and fostering continuous innovation alongside AI development requires transcending boundaries, including national borders, industries, and demographics. Emphasizing global collaboration, diversity, and industry-academia-government cooperation is vital for progress. Equal collaboration among universities, research institutions, and companies, with fluid human resource movement, is essential. Efficient and safe AI implementation requires methods to confirm quality, reliable AI, and effective data collection. Establishing AI engineering, ethical considerations, and economic aspects is crucial. Privacy-focused platforms enabling cross-border data utilization are needed, supported by shared computer resources and high-speed networks. Regulatory reforms are imperative across sectors to ensure an efficient and beneficial society driven by AI technologies. |

Governance Guidelines for Implementation of AI Principles (January 2022)

Japan's Ministry of Economy, Trade, and Industry (METI)'s Study Group on the Implementation of the AI Principles released the "Governance Guidelines for Implementation of AI Principles" (AI 原則実践のためのガバナンス・ガイドライン) (Ver. 1.1)²² for public comments on 28 January 2022.

These Guidelines, which are not legally binding, outline an approach for "**AI Businesses**" that are involved in the life cycle of AI systems to implement the Social Principles of Human-Centric AI within their organizations. AI Businesses include:

- » Entities that develop AI systems, whether for their own use or to provide the system to other businesses (developers).
- » Entities that operate AI systems, whether for their own use or for the use of others as a business (operators).

- » Entities that simply use an AI system developed by a developer or provided by an operator, and that is not responsible for the operation of the AI system and/or maintenance of its performance (users).
- » Entities that, as a business, provides others with data collected from a number of unspecified sources, data collected from specified people, data prepared by the data provider itself; a combination of them; or data created by processing the above-mentioned data, for the purpose of AI system training (data providers),

This approach is based on "**action targets**" for establishing an internal "AI Management System." Each action target is supported by examples of implementation methods, drawn from feedback from industry. While the action targets are intended to be sufficiently general and objective as to apply to all AI Businesses, the Guidelines leave it to each business to decide whether to adopt the examples of specific implementation measures and whether to do so in whole or in part.

| | |
|------------------------------|---|
| Conditions and Risk Analysis | <p>1-1 Understanding positive and negative impacts of using AI.</p> <p>Developers and operators should understand not only positive impacts but also negative impacts that AI systems may have, including unintended risks.</p> <p>This information should be reported to the top management and shared among those in top managerial positions, and their understanding should be updated in a timely manner.</p> |
| | <p>1-2 Understanding social acceptance of the use of AI.</p> <p>Before full-scale provision of the AI systems, Developers and operators should understand the current state of social acceptance based on opinions of not only direct stakeholders, but potential stakeholders.</p> <p>In addition, even after the full-scale operation, companies should obtain opinions of stakeholders again and update their perspectives in a timely manner.</p> |
| | <p>1-3 Understanding the company's AI proficiency.</p> <p>Developers and operators should evaluate and re-evaluate in a timely manner their AI proficiency based on:</p> <ul style="list-style-type: none"> » the extent of the company's experience in developing and operating AI systems; » the number of employees, including engineers, involved in the development and operation of AI systems and their degree of experience; and » the degree of AI literacy of these employees with respect to AI technology and ethics, <p>except in situations where a company assesses that negative impacts of their AI system are minor.</p> <p>If the negative impacts are assessed to be minor and no evaluation of AI proficiency is carried out, companies should be prepared to explain their rationale to their stakeholders.</p> |

| Stage | Action Targets |
|--|--|
| Goal Setting | <p>2-1 Considering and setting AI governance goals.</p> <p>Developers and operators should consider whether or not to set their own AI governance goals based on the Social Principles of Human-Centric AI.</p> <p>If a company decides not to set AI governance goals based on the assessment that their potential negative impacts are minor, they should be prepared to explain their rationale to their stakeholders.</p> |
| <p>Designing an AI Management System to Achieve AI Governance Goals</p> <p>This includes both technological and organizational systems.</p> | <p>3-1 Employing “gap analysis” between AI governance goals and the current state of AI governance and addressing gaps.</p> <p>Developers and operators should identify a gap between AI governance goals and current state in their AI systems and evaluate the impacts of these gaps as a starting point for improvement.</p> <p>Companies should provide sufficient information about the gaps and measures to address the gaps, as well as make a contact point easily accessible.</p> <p>To ensure that developers can appropriately conduct gap analysis, data providers should provide information on the data sets including data collection sources, collection policies, collection criteria, annotation criteria, and limitations on use.</p> <p>Developers should acquire data sets from data providers that provide sufficient information.</p> <p>3-2 Improving the literacy of AI management personnel.</p> <p>Developers and operators should strategically improve their AI literacy in order to properly operate their AI management system, considering outside learning materials as an option.</p> <p>Data providers should take steps to improve their employees’ general literacy in AI ethics by referring to practical examples for AI system developers and operators.</p> <p>3-3 Reinforcing AI management through cooperation between companies.</p> <p>Developers and operators and data providers should clarify and actively share AI system operational issues that the company or department is unable to fully address on their own and the information necessary to address these issues.</p> <p>AI system developers, operators, and data providers are encouraged to agree on scope of information disclosure in advance and consider measures to protect trade secrets, for example, by entering a non-disclosure agreement.</p> <p>Developers and operators should regularly collect relevant information, such as formulation of rules for the development and operation of AI systems, best practice, and incidents, and encourage the exchange of views within and outside the company.</p> <p>3-4 Preventing and responding to incidents.</p> <p>Developers and operators and those that provide data should, under the leadership of top management, reduce incident-related burdens on users by preventing incidents and through early response.</p> <p>They should consider defining response guidelines and plans so they can promptly notify users of AI incidents or disputes. They should identify the extent of the impact and damage, clarify legal responsibilities, consider relief measures and measures to prevent the spread of damage and recurrence, or take other relevant actions.</p> <p>Further, they should consider conducting rehearsal exercises relevant to such guidelines and plans, as appropriate.</p> |

| Stage | Action Targets |
|------------------------------------|---|
| Implementation | 4-1 Ensuring readiness to explain the implementation status of the AI management system. Developers and operators should make sure that they are ready for explanation about the implementation status of AI management systems externally by recording the gap analysis process under Action Target 3-1 and by taking other relevant actions. |
| | 4-2 Ensuring readiness to explain the operating status of individual AI systems. Developers and operators should monitor and record the status of preliminary and full-scale operations so that gap analysis for individual AI systems in preliminary and full-scale operations can be continuously implemented. Companies that develop AI systems should assist the monitoring conducted by companies that operate AI systems. |
| | 4-3 Considering proactively disclosing information on AI governance, including through the organization's Corporate Governance Code. Developers and operators should consider providing information relevant to AI governance as non-financial information in their Corporate Governance Codes and proactively disclosing such information. Non-listed companies should also consider proactively disclosing information related to AI governance activities. If companies decide not to disclose such information after due consideration, they should be prepared to explain the reason externally. |
| Evaluation | 5-1 Verifying an AI management system works appropriately. Individuals independent of the design and operation of the AI management system should verify whether an AI management system (e.g., a gap analysis process) is appropriately designed and operated for the achievement of the AI governance goals. |
| | 5-2 Considering seeking feedback from external stakeholders. Developers and operators should consider seeking opinions on their AI management system and the implementation of such a system from not only their shareholders but also from various stakeholders. If companies decide not to seek opinions outside after due consideration, they should be prepared to explain the reason externally. |
| Re-analysis of Conditions and Risk | 6-1 Re-implementing Action Targets 1-1 to 1-3 in a timely manner. Developers and operators should conduct re-evaluations, update their understanding, obtain new points of view, or take other relevant actions with respect to Action Targets 1-1 through 1.3, in a timely manner. |

Personal Information Protection Commission's Notices (June 2023)

On 2 June 2023, Japan's Personal Information Protection Commission (PPC) issued two guidance documents on the measures to implement under Japan's data protection law when using generative AI services. These documents include: (1) a **"Notice Regarding Cautionary Measures on the Use of Generative AI Services"** which outlines general guidance on the use of generative AI services; and (2) a **"Cautionary Notice"** to Open AI, which outlines specific guidance for OpenAI regarding ChatGPT's collection and use of sensitive personal data.²³

Both documents are intended to be non-exhaustive. The PPC acknowledges that its guidance is based on a point-in-time assessment of the data protection issues arising from the use of generative AI and highlights that it may take such additional measures as are necessary to respond to new developments in the technology.

The **Notice Regarding Cautionary Measures on the Use of Generative AI Services** recommends measures for three kinds of organizations: (1) businesses; (2) administrative agencies; and (3) general users to implement when using generative AI services.

Regarding **businesses**, the Notice recommends that businesses should observe the principle of **purpose limitation** when disclosing personal data to generative AI services. They should only include personal data in a prompt to a generative AI service if doing so is necessary to achieve the purpose for processing the personal data that the business has clearly identified and has notified to the data subject. If such disclosure does not fall within this specified purpose, then the Notice recommends that businesses should obtain **consent** from the data subject.

The Notice also recommends that before entering personal data in a prompt to a generative AI service, the business should confirm that the service provider does not retain the data and use it to further train the AI model. The Notice highlights the risks that when generative AI models are trained on such data, they may generate output that is **inaccurate**.

Similar guidance is provided to **administrative agencies**.

For **general users of generative AI services**, the Notice reiterates the risk that generative AI services may produce inaccurate output and recommends that users thoroughly review service providers' terms of use and privacy policies before using their services.

The Cautionary Notice to OpenAI contains two substantive recommendations for OpenAI regarding ChatGPT.

Firstly, the Notice highlights the need for a **legal basis** (such as consent) to collect sensitive personal data from users and other individuals and recommends the following:

- » Implementing the principle of **data minimization**: avoiding collecting sensitive personal data and taking measures to minimize the presence of sensitive personal data in any information collected from users.
- » Promptly **deleting** sensitive personal data or **anonymizing** it before using it to train a generative AI model.
- » Establishing a mechanism to comply with **requests from individuals for deletion of their sensitive personal data** where such data has already been used to train the model, unless there are legitimate reasons for refusing such requests.

- » Enabling users of ChatGPT to **opt-out** of use of information from users' prompts to further train the AI model.

Secondly, the Notice highlights the need to inform users and other parties of the purpose(s) for which ChatGPT collects and uses personal data. The Notice emphasizes that such information should be provided in Japanese.

Guidelines for AI Business Operators (April 2024)

On 21 December 2023, Japan's Ministry of Internal Affairs (MIC) and Ministry of Economy, Trade and Industry (METI) released an initial draft of its "**Guidelines for AI Business Operators**" for public consultation until 19 February 2024.²⁴ METI released the final version of the Guidelines in April 2024.²⁵

The Guidelines aim to unify and update Japan's voluntary AI governance framework, especially in response to the emergence of "**advanced AI systems**," such as foundational models and generative AI.

The Guidelines are based on the same fundamental principles as those in the Social Principles for Human-Centric AI, and the agile governance model recommended in METI's AI Governance Guidelines.

The Guidelines apply to all forms of AI and all organizations, whether in the private or public sector, that use AI in business activities. They provide recommendations across the lifecycle of an AI system for the following actors.

- » businesses that develop AI systems (**developers**).
- » businesses that provide services incorporating AI systems to business users and are responsible for operating such services or providing operational support (**providers**).
- » businesses that use AI systems or services in their business activities (**business users**).

Recommendations for **developers** at different stages of the AI cycle include:

| Stage | Recommendation |
|---|---|
| Data pre-processing and training | Implementing “Privacy by Design” principles to ensure that personal data is collected appropriately. |
| | Complying with laws and regulations governing protection of personal data, intellectual property, and confidential information. |
| | Implementing a system to manage access to data before and during training. |
| | Taking reasonable measures to control the quality of training data, and conducting parallel development to minimize bias. |
| AI development | Ensure that the AI system can maintain its performance level under various conditions, not just the expected usage conditions. |
| | Implementing appropriate safety measures to minimize risks of harm to stakeholders. |
| | Considering the possibility that bias may be introduced through each technical component of the AI model, and conducting parallel development to minimize bias. |
| | Where relevant, selecting only an appropriate pre-trained model for fine-tuning. |
| | Ensuring verifiability, including by maintaining records for post-verification. |
| After AI development | Remaining informed of cybersecurity trends and emerging cyber threats. |
| | Providing information to relevant stakeholders (including through AI providers) on the AI system, including: <ul style="list-style-type: none"> » the possibility of changes in the output or program due to AI system training; » technical characteristics of the AI system, mechanisms for ensuring safety, foreseeable risks that may arise from its use, and mitigation measures; » the intended scope of use by AI developers; » the operational status of the AI system, the cause of malfunctions, and the response status; » the content and reasons for AI updates; » data collection policies, learning methods, and implementation systems for data used to train the AI model. |
| | Informing and explaining to AI providers that AI systems may experience significant changes in predictive performance and output quality, or may not reach the expected accuracy after deployment, and the resulting risks. |
| | Documenting the AI system development process, data collection and labeling that influence decision-making, and algorithms used, in a way that allows third-party verification as much as possible. |
| | Contributing to the creation of innovation opportunities. |

Recommendations for **providers** at different stages of the AI cycle include:

| Stage | Recommendation |
|--|--|
| Implementation of the AI system | Ensure that the AI system can maintain its performance level under various conditions, not just the expected usage conditions. |
| | Implementing appropriate safety measures to minimize risks of harm to stakeholders. |
| | Use the AI appropriately within the scope set by the AI developer. Consider whether there are any differences between the assumed usage environment set by the AI developer and the actual usage environment. |
| | Ensure the fairness of the data and consider the biases in the information referenced or external services connected. |
| | Periodically evaluate the input, output, and reasoning of the AI model, and monitor for the occurrence of biases. If necessary, request the AI developer to re-evaluate the biases in the various technical components of the AI model and provide feedback on the evaluation results to drive improvements to the AI model. |
| | Consider the possibility of biases being introduced in the AI system/service or user interface that receives the AI model's output, which could arbitrarily constrain business processes or the judgments of AI users or non-users. |
| | Implement appropriate privacy protection and security measures. |
| | Document the system architecture and data processing flow of the provided AI system/service that influence decision-making. |
| After Providing the AI System/Service | Periodically verify that the AI system/service is being used for appropriate purposes. |
| | Gather information on privacy infringements in the AI system/service, appropriately address any incidents, and consider measures to prevent recurrence. |
| | Take note of emerging attack methods against AI systems/services and consider resolving vulnerabilities. |
| | Promptly provide information on the provided AI system/service, in a simple and accessible form, such as: <ul style="list-style-type: none"> » The fact that AI is being used and appropriate/inappropriate usage methods. » the possibility of changes in the output or program due to AI system training; » technical characteristics of the AI system, mechanisms for ensuring safety, foreseeable risks that may arise from its use, and mitigation measures; » the operational status of the AI system, the cause of malfunctions, and the response status; » the content and reasons for AI updates; » data collection policies, learning methods, and implementation systems for data used to train the AI model. |
| | Encourage appropriate use by AI users and provide them with the following information: <ul style="list-style-type: none"> » Reminders about using data with assured accuracy and, if necessary, timeliness. » Warnings about the risk of inappropriate AI model learning through context-based learning. » Precautions when inputting personal information. » Warn about inappropriate input of personal information to the provided AI system/service. |
| | Prepare service terms and conditions for AI users and non-users. |
| | Clearly state the privacy policy. |

Recommendations for **business users** include:

| Stage | Recommendation |
|---|--|
| When using an AI system or service | Use the AI system/service within the range designed by the AI provider, in compliance with the usage precautions defined by the AI provider. |
| | Ensure that data is entered in an accurate and if necessary, timely manner. |
| | Understand the accuracy and risk level of the AI output, and use it after confirming various risk factors. |
| | Ensure that data is input fairly to avoid significant unfairness, and make responsible judgments on the business use of AI output results, bearing in mind potential bias. |
| | Be careful not to inappropriately input personal information into the AI system/service. |
| | Gather information on privacy infringements in the AI system/service and consider preventive measures. |
| | Comply with the security precautions provided by the AI provider. |
| | Obtain output results from the AI system/service by inputting data with assured fairness and being mindful of biases in the prompts. When utilizing the output results for business decisions, inform the relevant stakeholders. |
| | Provide information, including on appropriate usage methods, to relevant stakeholders in a simple and accessible form, to a reasonable extent. |
| | If the business user plans to use data provided by relevant stakeholders, inform them in advance about the characteristics and applications of the AI, the contact points with the provider, the privacy policy, and the means and format of data provision. |
| | Set up a point of contact to respond to inquiries from relevant stakeholders, in collaboration with the AI provider. |
| | Properly store and utilize the documents provided by the AI provider on the AI system/service. |
| | Comply with the service terms and conditions defined by the AI provider. |

The Draft Guidelines also encourage organizations to comply with relevant obligations under: (1) the International Guiding Principles for Organizations Developing Advanced AI Systems; and (2) International Code of Conduct for Organizations Developing Advanced AI Systems, both of which were released in October 2023 under the G7 Hiroshima AI process.

Annex A to the Draft Guidelines provides a non-exhaustive a number of potential risks arising from AI, including generative AI, based on a review of existing cases:

- » **Biased or discriminatory outputs.**
- » **Creation of “filter bubbles” and amplification of bias.**
- » **Loss of diversity** in content and opinions.
- » **Inappropriate handling of personal data**, including lack of transparency in use of personal data, and use or disclosure of personal data without data subject’s knowledge or consent, and

- » **Harm to individuals’ physical and mental wellbeing and property.**
- » **Cyberattacks and jailbreaking of AI systems for malicious use.**
- » **Environmental impact.**
- » **Fraud.**
- » **Breaches of personal data or confidential information.**
- » **Factual inaccuracies**, which individuals may rely on to their detriment.
- » **Spreading misinformation and disinformation.**
- » **Intellectual property infringement.**
- » **Breaches of laws and regulations governing professions**, such as law and medicine.

Singapore

Singapore's approach to governance of generative AI has been led by the Infocomm Media Development Authority (IMDA) and represents a proactive and collaborative effort to develop governance frameworks specifically tailored for the unique challenges posed by generative AI technologies.

This inclusive process allows for comprehensive consideration of technical, ethical, and legal dimensions to inform robust governance mechanisms appropriate for this powerful new technology domain.

By engaging a wide range of stakeholders including industry, researchers, and the public both domestically and internationally, Singapore aims to foster a trusted ecosystem that facilitates innovation while mitigating risks.

Model AI Governance Framework (January 2020)

On 23 January 2019, Singapore's Infocomm Media Development Authority (IMDA) released the first edition of its **"Model AI Governance Framework."** A second edition of the Model Framework was released on 21 January 2020.²⁶

The Model AI Governance Framework defines AI as "a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning, and, depending on the AI model, produce an output or decision (such as a prediction, recommendation, and/or classification)."

It outlines practical guidance for private sector organizations that deploy AI at scale to:

- » build stakeholder confidence in AI by enabling organizations to use AI responsibly and manage risks throughout deployment of AI; and
- » demonstrate reasonable efforts to align their internal policies, structures, and processes with relevant accountability-based practices in data management and protection.

This guidance is non-binding – the Framework is meant to be flexible and permits organizations to adopt such recommendations as are relevant to them, and adapt these recommendations to suit their needs.

The Framework identifies the following actors in the AI value chain:

- » **"AI Solutions Providers"** – i.e., developers of AI solutions or applications that make use of AI technology; device manufacturers that integrate AI-powered features into their products; and developers of solutions that are not standalone products but are meant to be integrated into a final product.
- » **"Organizations"** – i.e., companies or entities that adopt or deploy AI solutions in their operations.
- » **"Individuals"** – i.e., the persons to whom organizations intend to supply AI products and services.

The framers of the Framework made a conscious decision not to articulate a new set of ethical principles for AI. Instead, the Framework sets out practical considerations guiding organizations to deploy AI responsibly, based on commonly accepted ethical principles. That said, the Framework expressly states that it is based on two fundamental principles:

- » In order to build trust and confidence in AI, AI-based decision-making should be **explainable, transparent, and fair**.
- » AI solutions should be **human-centric** (e.g., amplifying human capabilities and protecting the interests of human beings, such as their wellbeing and safety).

A longer list of 12 AI ethics principles is presented in Appendix A to the Framework as a glossary for organizations seeking to develop their own internal AI policies; however, not all of these principles are directly addressed by the Model AI Governance Framework.

The Framework is organized into 4 areas of a generalized AI deployment lifecycle:

| Section | Subsections |
|---|--|
| Internal governance structures and measures | <p>Assignment of roles and responsibilities within an organization. The Framework encourages organizations to allocate responsibility for and oversight of the various stages and activities in AI deployment to appropriate departments and personnel.</p> <p>Examples of roles and responsibilities that could be allocated include:</p> <ul style="list-style-type: none"> » Assessing and managing the risks of deploying AI; » Maintaining, monitoring, document, and reviewing AI models that have been deployed; » Provide effective feedback and disclosure channels to stakeholders; » Training personnel to interpret the AI model and work with the AI system. |
| | <p>Standard operating procedures for monitoring and managing risk. The Framework recommends that organizations consider implementing a risk management system and internal controls that specifically address the risks involved in the deployment of the selected AI model.</p> <p>Examples of possible measures that could be implemented include:</p> <ul style="list-style-type: none"> » Ensuring that the datasets used to train AI models are adequate for the intended purpose; » Assessing and managing the risks of inaccuracy or bias during model training; » Establishing monitoring and reporting systems, with appropriate channels to management. » Reviewing internal governance structures and measures and ensuring proper knowledge transfer whenever there are changes in key personnel involved in AI activities. » Periodically reviewing the internal governance structure and measures to ensure their continued relevance and effectiveness. |
| Human involvement in AI-augmented decision making | <p>The Framework outlines guidance to help organizations determine the appropriate level of human involvement in AI-augmented decision-making, based on a risk impact assessment.</p> <p>It outlines a spectrum of approaches, from human-in-the-loop, to human-out-of-the-loop, to human-over-the-loop and provides examples of when each may be appropriate, taking into account the probability and severity of potential harms.</p> |
| Operations management | <p>The Framework outlines good data accountability practices for training datasets used to train AI models. These include:</p> <ul style="list-style-type: none"> » Understanding the lineage of the data and maintaining data provenance records; » Ensuring the quality of the data based on factors like accuracy, completeness, recency, relevance; » Identifying and addressing biases inherent in the data; » Considering the use of different datasets for training, testing, and validation; and » Periodically reviewing and updating datasets. |
| | <p>The Framework also outlines possible measures to ensure that the AI model makes decisions that are explainable, repeatable, robust, reproducible, and auditable.</p> <p>These include assessment and testing, as well as regular model tuning to respond to changes over time.</p> |

| Section | Subsections |
|---|--|
| Stakeholder interaction and communication | <p>The Framework discusses different strategies for organizations to build trust with stakeholders.</p> <p>These include:</p> <ul style="list-style-type: none"> » Publishing general information on their use of AI, how the AI model makes decisions, and the organization's policies in relation to AI; » Developing policies on what explanations to provide to individuals, and when. <p>It also discusses relevant factors, such as audience, purpose, and context.</p> |
| | <p>The Framework discusses potential measures that organizations could consider implementing to manage relationships with their customers, such as:</p> <ul style="list-style-type: none"> » providing opt-outs; » Creating channels for customers to provide feedback or raise queries; » Establishing mechanism for customers to request a review of AI decisions that have affected them materially; and » Providing acceptable use policies. |

Annex B to the Framework outlines measures for auditing algorithms.

The Framework is accompanied by two other guidance documents: (1) the Implementation and Self Assessment Guide for Organisations (ISAGO); and (2) a Compendium of Use Cases (Compendium), split into two volumes.

- » The ISAGO strives to assist organizations in evaluating the compatibility of their AI governance practices with the Model Framework. Additionally, it offers a comprehensive collection of valuable industry examples and practices to aid organizations in the implementation of the Model Framework.
- » The Compendium, comprising two volumes, showcases how organizations, both local and international, across various sectors and sizes, have implemented or harmonized their AI governance practices with all sections of the Model Framework. The Compendium also highlights how these featured organizations have successfully established accountable AI governance practices and derived benefits from the incorporation of AI into their business operations.

As it was released roughly 4 years before the public launch of ChatGPT, the Framework was not written with present-day generative AI systems in mind. However, earlier forms of generative AI technology appear to have been considered in drafting the Framework, as the Framework refers to the GPT-2 as a “next-generation AI powered natural text generator” capable of generating text that is difficult to distinguish from human-produced text.

Discussion Paper on Generative AI: Implications for Trust and Governance (June 2023)

On 7 June 2023, Singapore's Infocomm Media Development Authority (IMDA) and Aicadium, a Singapore-based AI company, published a discussion paper titled “**Generative AI: Implications for Trust and Governance**.”²⁷

The publication of this paper coincided with the launch of the **AI Verify Foundation**, a not-for-profit subsidiary of the IMDA intended to work with industry to support open-source development of the IMDA's AI testing framework, known as AI Verify.²⁸ Note that AI Verify currently does not apply to generative AI systems, including LLMs.²⁹

The paper outlines proposals for senior policymakers and business leaders on building an ecosystem for the trusted and responsible adoption of generative AI globally and invites comments from global stakeholders.

The paper begins by providing a brief overview of generative AI technology, as well as opportunities and challenges. **Challenges** highlighted in the paper include:

- » Factual inaccuracies.
- » Leaking personal data or confidential information.
- » Scaling disinformation, toxicity, and cyber-threats.
- » Challenges for intellectual property law.
- » **Bias.**
- » **Ensuring that generative AI aligns with human values and goals.**

Later sections of the paper identify six core areas that make up a proposed approach to governance of generative AI that builds on existing frameworks: (1) **accountability**; (2) **data**; (3) **model development and deployment**; (4) **assurance and evaluation**; (5) **safety and alignment research**; and (6) **“Generative AI for Public Good.”** The paper then recommends governance measures that could be adopted to enhance trust and safety within each of these areas:

| Area | Proposed Measures |
|---|---|
| Accountability | The paper recommends that adopting a shared responsibility framework (similar to that adopted by major cloud service providers) among the different parties involved in the life cycle of generative AI systems could clarify responsibilities and incentivize safer outcomes. |
| | The paper recommends that developers could be required to provide information about generative AI models in a standardized format (similar to “nutrition labels”). The paper suggests that such information would help deployers to make proper risk assessments. |
| | The paper recommends that labeling or watermarking of AI-generated content could allow consumers to make more informed decisions and choices, and allow content distributors to take remedial actions to prevent the distribution of harmful content. |
| Data | The paper recommends that developers should be transparent about the types of datasets used to train generative AI models . |
| | The paper recommends that policymakers should provide guidance on the requirements for data privacy and copyright under their respective regulations. |
| | The paper recommends that stakeholders consider collaborating on building trusted data repositories that generative AI models could reference to mitigate bias embedded in their training datasets. |
| Model development and deployment | The paper recommends that developers should be transparent about how their models are developed and tested . |
| | The paper recommends that developers and deployers should partner on monitoring the performance of generative AI models . |
| | The paper suggests that policymakers can support developers and deployers by facilitating the development of standardized metrics and tools to evaluate model safety, performance, efficiency, and environmental sustainability. |
| | The paper recommends that policymakers also carefully deliberate their approach to regulating AI and adopt a calibrated approach , using or updating existing laws as necessary. |
| Assurance and evaluation | The paper suggests that there may be value in independent third-party evaluation and assurance to provide objective assessments. |
| | The paper also suggests that development of evaluation and assurance tools and testing of AI models would benefit from involvement from the open-source community . |
| Safety and alignment research | The paper recommends that policymakers invest in safety alignment and research to enable interpretability, controllability, and robustness of generative AI systems. |
| Generative AI for Public Good | The paper recommends the creation of consumer literacy programs to promote public understanding and safe use of generative AI. |
| | The paper also recommends providing greater education and training on generative AI-related skills to address changes to work from adoption of generative AI. |
| | The paper recommends that policymakers update their guidance to make generative AI accessible to all enterprises, including providing examples of use cases. |
| | The paper recommends that policymakers also consider establishing common infrastructure that the wider ecosystem can use to develop and test generative AI models and applications. |
| | The paper recommends that stakeholders assess the impact of generative AI on end-users and develop measures to quantify such impact. |
| | Lastly, the paper calls for international collaboration on generative AI governance, bringing together diverse stakeholders. |

Generative AI Sandbox and Draft Catalogue of LLM Evaluations (October 2023)

On 31 October 2023, the IMDA and the AI Verify Foundation announced the launch of a regulatory sandbox for the evaluation of generative AI featuring several major domestic and multinational companies.³⁰

To guide the sandbox, the IMDA also released a draft catalog of current benchmarks and methods to evaluate LLMs, titled “**Cataloguing LLM Evaluations**,”³¹ for public comment. The catalog is divided into three parts.

- » **Part 1** compiles commonly used technical testing tools organized into 5 categories reflecting what these tools test for, as well as their methods: (1) General Capabilities; (2) Domain Specific Capabilities, subcategorized into (a) law, (b) medicine, and (c) finance; (3) Safety and Trustworthiness; (4) Extreme Risks; and (5) Undesirable Use Cases.
- » **Part 2** analyzes the LLM evaluation landscape, highlighting key areas for further development, such as the need for more context-specific evaluations, frontier model evaluations and the need for standards and best practices.
- » **Part 3** recommends a baseline set of evaluation tests for use in generative AI products. These evaluations comprise 5 attributes that LLMs should be tested on pre-deployment to ensure a minimal level of safety and trustworthiness: (1) **bias**; (2) **factuality**; (3) **toxicity generation**; (4) **robustness**; and (5) **data governance**.

Proposed Model AI Governance Framework for Generative AI (January 2024)

On 16 January 2024, the IMDA and the AI Verify Foundation released the “**Proposed Model AI Governance Framework for Generative AI**”³² for public comment.

While the Proposed Framework adopts a similar title to the IMDA’s existing Model AI Governance Framework (see above), the Proposed Framework follows a different approach. Whereas the Model AI Governance framework was intended to provide guidance to organizations that had decided to deploy AI technologies at scale, the Proposed Framework proposes a broader approach that:

- » aims to build a trusted ecosystem for generative AI, addressing new concerns while continuing to facilitate innovation;
- » involves all key stakeholders, including policymakers, industry, the research community, and the broader public, internationally; and
- » emphasizes the need to review existing governance frameworks.

In this regard, the Proposed Framework’s approach builds on recommendations made in the earlier Discussion Paper. In particular, it refines the core areas proposed in the Discussion Paper and adds three new areas (new additions are underlined): (1) **accountability**; (2) **data**; (3) **trusted development and deployment**; (4) **incident reporting**; (5) **testing and assurance** (formerly, assurance and evaluation); (6) **security**; (7) **content provenance**; (8) **safety and alignment research and development**; and (9) “**AI for Public Good**.”

The Proposed Framework also highlights several regulatory actions and governance measures that various stakeholders could consider adopting to enhance trust and safety within these areas:

| Area | Proposed Measures |
|---|---|
| Accountability | <p>The Proposed Framework suggests that stakeholders should consider allocating responsibility to end-users on both an ex-ante (addressing risks before they arise) and ex-post (addressing risks after they arise) basis.</p> <p>For ex-ante allocation, the Proposed Framework suggests that responsibility should be allocated based on the level of control that each stakeholder has in the generative AI life cycle. It repeats the Discussion Paper's suggestion that generative AI governance could adopt a shared responsibility model like that currently employed by several cloud service providers.</p> <p>The Proposed Framework specifically recommends that developers of AI models could lead development of trusted platforms for deployers to obtain AI models to avoid the risk that models are tampered with.</p> <p>For ex-post allocation, the Proposed Framework highlights the challenges in allocating responsibility for new and unanticipated issues and calls on policymakers to consider updating their legal frameworks.</p> |
| Data | <p>The Proposed Framework calls on policymakers to engage in dialogue with stakeholders and issue guidance on the use of data in model development, particularly around the application of existing data protection and intellectual property laws.</p> <p>Data protection issues highlighted in the Proposed Framework include the legality of web-scraped datasets, legal bases for processing personal data, and the role of Privacy Enhancing Technologies, including anonymization.</p> <p>The Proposed Framework recommends that developers undertake data quality control measures and adopt general best practices in data governance, including annotating training datasets consistently and accurately, and using data analysis tools to facilitate data cleaning.</p> <p>The Proposed Framework also suggests that stakeholders should consider expanding the available pool of trusted reference datasets for model development, benchmarking, and evaluation.</p> <p>It highlights that governments could play a role in curating repositories of representative training data sets for their specific cultural, social, or linguistic contexts.</p> |
| Trusted Development and Deployment | <p>The Proposed Framework stresses the need for baseline safety practices and highlights several practices on which industry appears to align, including risk assessments, fine-tuning techniques such as Reinforcement Learning from Human Feedback, user interaction techniques such as input and output filters, and techniques like Retrieval-Augmented Generation and few-shot learning to reduce hallucinations and improve accuracy.</p> <p>The Proposed Framework repeats the Discussion Paper's recommendation for standardized disclosure mechanisms for AI models and elaborates on potential areas that these mechanisms could cover.</p> <p>It also stresses the need for greater transparency to governments for models that pose potentially high risks, such as advanced models that have national security or societal implications.</p> <p>The Proposed Framework also calls for a comprehensive, systematic approach to safety evaluation and highlights that additional evaluations may be needed for certain sectors or domains.</p> <p>It recommends that industry and sectoral policy makers jointly improve evaluation benchmarks and tools, while still maintaining coherence between baseline and sector specific requirements.</p> |

| Area | Proposed Measures |
|--|---|
| Incident Reporting | The Proposed Framework calls for stakeholders to establish structures and processes to report cybersecurity vulnerabilities and incidents in relation to generative AI models and systems. |
| Testing and Assurance | Like the earlier Discussion Paper, the Proposed Framework suggests that third-party testing and assurance could play a useful role in the generative AI ecosystem. It suggests that stakeholders could draw from existing audit practices but should develop standardized benchmarks and methodologies. It also suggests creating accreditation mechanisms. |
| Security | The Proposed Framework highlights that AI security is a nascent field so stresses the importance of implementing a “ Security by Design ” approach and developing new safeguards , such as input filters to detect unsafe prompts, and digital forensic tools for generative AI. |
| Content Provenance | The Proposed Framework highlights challenges from highly realistic synthetic content , the need for technical solutions , such as digital watermarking and cryptographic provenance, to show that content was generated or modified by AI, and policies to support these solutions. |
| Safety and Alignment Research and Development | The Proposed Framework highlights the need for human capabilities to align and control AI to keep pace with advancements in AI models. This entails greater international coordination in research and development of model safety and alignment. |
| AI for Public Good | Building on the Discussion Paper’s recommendations, the Proposed Framework identifies 4 areas that could help to ensure that AI brings long-term benefits: <ul style="list-style-type: none"> » Democratizing access to technology, through human-centric design, digital literacy initiatives, and public-private initiatives to drive innovation and use of AI by small- and medium-sized enterprises. » Delivery of public services using AI. » Upskilling the workforce and redesigning jobs. » Sustainability. |

South Korea

South Korea has been working towards comprehensive national AI legislation since 2021. Progress has been limited since.

In the interim, South Korea’s data protection authority (PIPC) has been proactive in establishing sector-specific governance and pursuing enforcement action against OpenAI, including fining OpenAI in July 2023 for infringing South Korea’s data protection law over ChatGPT’s handling of personal data.

Human Centered AI Ethics Standards (December 2020)

South Korea’s “**Human Centered AI Ethics Standards**” (사람이 중심이 되는 인공지능 윤리기준) (HCAIE Standards)³³ were released on December 23, 2020, at the 19th meeting of the Presidential Committee on the 4th Industrial Revolution.³⁴

The HCAIE Standards were the product of several South Korean government agencies, including the Ministry of Science and ICT (MSIT) and the Korea Information Society Development Institute.³⁵

The HCAIE Standards are voluntary and intended to serve as a reference point for all members of society – including government, the public and private sectors, and the public – to realize “human-centered AI” throughout the AI lifecycle.

They provide a flexible set of general principles that are not limited to any specific domain, issue, or technology. The Standards drew inspiration from the OECD AI Principles as well as other regional frameworks, such as Japan’s Social Principles.

The HCAIE Standards are structured into a hierarchy comprising: (1) three Basic Principles for human-AI relations to achieve “Human Centered AI;” and (2) 10 Requirements that give effect to the Basic Principles.

The three **Basic Principles** for human-AI relations to achieve “Human Centered AI” are:

| Basic Principle | Elaboration |
|--|---|
| Human Dignity (인간 존엄성 원칙) | Human beings have an intrinsic value that cannot be exchanged for a mechanical product, including AI. AI should be developed and used in a way that does not harm human life and mental and physical health. AI should be used and developed in a way that is safe and robust and that does not harm human beings. |
| Common Good of Society (사회의 공공선 원칙) | Society pursues the wellbeing and happiness of as many people as possible. AI should be developed and used in a manner that ensures accessibility for socially disadvantaged and vulnerable groups that are prone to marginalization in an intelligent information society. Development and use of AI for public good should enhance the wellbeing of humanity from a societal, national, and ultimately, a global perspective. |
| Reasonableness of Technology (기술의 합목적성 원칙) | The development and use of AI technology should be ethical and in accordance with AI technology’s purpose as a tool to improve human life. The development and use of AI technology to improve human life and prosperity should be encouraged and promoted. |

The **10 Requirements** that give effect to the Basic Principles are:

| Requirement | Elaboration |
|--|---|
| Human Rights Guarantees (인권보장) | The development and use of AI should respect the rights equally granted to all humans and guarantee the rights specified in various democratic values and international human rights law. The development and use of AI must not infringe on human rights and freedoms. |
| Protection of Privacy (프라이버시 보호) | Individual privacy should be protected throughout the entire process of development and use of AI. Efforts should be made to minimize the misuse of personal data throughout the entire lifecycle of AI. |
| Respect for Diversity (다양성 존중) | At all stages, the development and use of AI should represent and reflect the diversity of users, including gender, age, disability, race, religion, and country. Bias and discrimination based on personal characteristics should be minimized. Commercialized AI should be applied fairly to everyone. AI technology and services should be made accessible to socially underprivileged and vulnerable groups. Effort should be made to distribute the benefits of AI evenly to all people, not to specific groups. |
| Non-Infringement (침해금지) | AI must not be used for the purpose of causing direct or indirect harm to humans. Efforts should be made to mitigate the risks and negative consequences that AI may cause. |
| Public Nature of AI (공공성) | AI should be used not only for the pursuit of personal happiness, but also for the promotion of social publicness and the common benefit of humanity. AI should be used to drive positive social change. Comprehensive education should be implemented to maximize the positive functions of AI and minimize the negative functions. |

| Requirement | Elaboration |
|------------------------------------|--|
| Joint Action (연대성) | <p>Solidarity should be maintained in relationships between various groups, and AI should be used with sufficient consideration for future generations.</p> <p>Fair opportunities should be ensured for diverse stakeholders throughout the entire lifecycle of AI.</p> <p>Efforts should be made for international cooperation in the development and utilization of ethical AI.</p> |
| Data Management (데이터 관리) | <p>Individual data, including personal data, should be used in line with its intended purpose, and not for any other purpose.</p> <p>Data quality and risk must be managed to minimize data bias throughout the entire process of data collection and use.</p> |
| Accountability (책임성) | <p>Efforts should be made to minimize damage that may occur by establishing a responsible entity in the process of developing and using AI.</p> <p>Responsibilities between AI design and developers, service providers, and users must be clearly specified.</p> |
| Safety (안전성) | <p>Efforts must be made to prevent potential risks and ensure safety throughout the entire process of developing and use of AI.</p> <p>Efforts should be made to ensure that users have the ability to control the operation of an AI when an obvious error or infringement occurs.</p> |
| Transparency (투명성) | <p>Efforts should be made to increase the transparency and explainability of AI for the purpose of building social trust, and to increase the transparency and explainability of AI, and take into account conflicts with other principles.</p> <p>Advance notice should be provided of significant considerations, such as the nature of AI use and potential risks when offering products or services based on AI.</p> |

The HCAIE Standards also outline future plans on the part of the South Korean government to promote these principles through education, development of metrics, and continuation of the discussion.

Bill on Fostering Artificial Intelligence and Creating a Foundation of Trust (July 2021)

On 1 July 2021, a draft AI law, titled the “**Bill on Fostering Artificial Intelligence and Creating a Foundation of Trust**,”³⁶ was introduced in the National Assembly, South Korea’s unicameral national legislature, by 23 National Assemblymen.

The Bill aims to contribute to the development of South Korea’s AI industry and if enacted, would lay the groundwork for further action by the South Korean Government.

In particular, it provides a statutory basis for the South Korean Government take several measures in relation to AI, including:

- » Enacting a set of binding “Ethical Principles for an AI Society” in the form of a Presidential Decree;
- » Enabling the Minister for Science and IT to establish and implement a Basic Plan on a tri-annual basis;

- » Establishing an AI Society Committee to deliberate on government AI plans;
- » Empowering the Minister for Science and IT to develop further AI policies, release AI standards, conduct investigations, and impose penalties.
- » Designating AI systems that may pose a risk to humans’ rights and interests as “AI Systems for Special Use” and subjecting them to additional reporting obligations.

Updates on the progress of the Bill since July 2021 then have been limited.

On 16 August 2023, South Korea’s MSIT announced that it had established an AI Legislation Committee to facilitate discussions on AI-related issues, as part of its comprehensive plan to create a roadmap for development of AI legislation.³⁷

PIPC Enforcement Decisions against OpenAI (July 2023)

In March 2023, the PIPC commenced an investigation into ChatGPT, based on reports that the service had leaked personal data.

On 27 July 2023, the PIPC announced the outcome of its investigation.³⁸ The PIPC imposed a fine of KRW 3.6 million (~US\$2,700) on OpenAI and identified

several areas in which the company had failed to comply with South Korea's Personal Information Protection Act (PIPA).

These included:

- » **Failing to report a breach of the personal data of 689 ChatGPT users in South Korea.** Notably, the PIPC did not find that OpenAI had failed to meet its obligations under the PIPA to secure the personal data. However, the PIPC still recommended measures to improve OpenAI's personal data processing systems to prevent recurrence of the issue.
- » **Failing to provide a privacy policy and consent procedure in Korean.**
- » **Failing to include certain information required by the PIPA in ChatGPT's privacy policy,** including specific methods and procedures for destroying personal data, lack of clarity as to OpenAI's agents in South Korea.
- » **Allowing minors to register for use of ChatGPT.** ChatGPT allowed users over the age of 13 to register for ChatGPT services, including consent to use of their personal data by the service. However, under the PIPA, only permits users over the age of 14 to give independent consent for processing of their personal data.

The PIPC also noted that OpenAI **was not sufficiently transparent** as to several matters that the PIPC considered were necessary to identify infringements of South Korean users' privacy. These included:

- » how ChatGPT collects and uses personal data;
- » the sources for its Korean-language training data;
- » its efforts to prevent ethical issues; and
- » methods for users to opt-out of collection of their personal data.

The PIPC gave OpenAI until 15 September 2023 to bring its processing of personal data into compliance with the PIPA.

Policy Direction for Safe Use of Personal Information in the AI Era (August 2023)

On 3 August 2023, the PIPC published its "**Policy Direction for Safe Use of Personal Information in the AI Era**" (Policy Direction).³⁹ The document outlines the PIPC's policy in relation to AI, focused on enabling the safe use of data for the development of AI systems while minimizing the risks of privacy infringement.

The Policy Direction aims to minimize the risk of privacy infringement from development and deployment of AI systems while allowing data that is essential for AI innovation to be used safely. In particular, it identifies the following risks:

- » **Privacy infringement.** The Policy Direction highlights that generative AI systems may process personal data in ways that data subjects may not expect and without establishing a relationship with data subjects, whether through consent or a contract (e.g., by processing personal data that has been scraped from the internet). It also highlights that generative AI has increased the scale in which privacy infringements like these occur.
- » **Identity threats.** The Policy Direction highlights that generative AI systems may produce factual inaccuracies and distortions that may threaten an individual's identity and undermine democratic values through the spread of misinformation. It also highlights that synthetic media such as "deepfakes" may be used for fraudulent purposes.

To address these new challenges, the Policy Direction outlines high-level guidance on data protection in the context of AI, including generative AI.

It starts by identifying the following data protection principles from the PIPA that are relevant to AI systems:

- » **Suitability for purpose:** The purpose for processing personal data should be clearly identified and explained and personal data should only be processed within the scope of that purpose, having regard to the data subject's rights and expectations.
- » **Lawful processing of personal data:** Personal data should be processed lawfully and justly, weighing the benefits that can be obtained from AI and the risks posed to the data subject.
- » **Accuracy, completeness, and currentness.** Personal data should be accurate, complete, and up-to-date. If there is an error or distortion of the data, the right to respond must be guaranteed.
- » **Transparency:** The data collection and processing methods of an AI system should be transparently disclosed.
- » **Safety management:** Continuous management system is required to ensure safety based on AI risk assessment.
- » **Guarantee of rights of data subjects' rights,** including the rights to correction, deletion, and suspension of processing, the right to refuse automated decisions, and the right to request an explanation.
- » **Minimizing privacy infringement:** Personal data should be processed in a way that minimizes the infringement of the data subjects' privacy.

The Policy Direction also provides guidance on data protection obligations and best practices at each stage of the life cycle of an AI system, from **development** (including **planning, data collection, and training**) and **deployment**.

| Stage | | Guidance |
|-------------|-----------------|---|
| Development | Planning | Implementing Privacy by Design principles : <ul style="list-style-type: none"> » Identifying relevant protections that apply at each stage of the AI life cycle. » Clarify the legal basis for collecting and using personal data. » Planning to respond to privacy issues. » Identifying and implementing measures to minimize errors and biases in the data. » Identifying and implementing measures to disclose important information, such as the source of training data and how personal data is processed. » Planning ways to guarantee data subjects' rights, and establishing and operating reporting channels. |
| | | Establishing a governance system , in which developers and data protection officers collaborate on analyzing risks and preparing strategies to mitigate them. |
| | Data Collection | Ensuring that there is a valid legal basis under the PIPA for collecting personal data and that the data is processed within the scope of the purpose for collection or for a purpose that is reasonably related to it. |
| | | Publicly available personal data may only be processed on the basis of consent or a legitimate interest or if it has been pseudonymized. |
| | | Complying with relevant PIPC guidelines for different types of personal data , such as visual image data, biometric data, etc. |
| | Training | Implementing protective measures, such as using Privacy Enhancing Technologies (PETs) including techniques like anonymization and pseudonymization , with safeguards against reidentification of data subjects. |
| Deployment | | Enhancing transparency by informing data subjects as to the purpose and method for collection and use of their personal data. |
| | | Enhancing explainability , for instance through model cards. |
| | | Implementing measures to prevent harms to data subjects , such as: <ul style="list-style-type: none"> » ensuring that the system refuses to generate responses to user prompts that induce inappropriate answers; » filtering of generated answers; and » establishing and operating AI risk management and response systems at all times. |
| | | Giving effect to data subjects' rights , including providing data subjects with clear, understandable information about how to exercise their rights. |
| | | |

In addition to the above guidance, the Policy Direction outlines several specific regulatory actions that the PIPC will take in future (see below).

| Category | Proposed Measures |
|--|--|
| Establishing a principle-based regulatory system to promote accountability and offer guidance on legal uncertainties | Establishing an AI Privacy Team within the PIPC to address AI-related matters. |
| | Introducing a regulatory sandbox, known as the Prior Adequacy Review System , that would enable the PIPC to exempt AI businesses from certain obligations under the PIPA. |

| Category | Proposed Measures |
|--|---|
| Preparing sectoral guidelines through public-private cooperation | Establishing an AI Privacy Public-Private Council to facilitate discussions between the public and private sectors and jointly develop guidelines for each sector. |
| | Expanding research and development on PETs and preparing guidelines on their use. |
| | Developing an AI risk assessment model , based on a regulatory sandbox, to allow regulations to be designed according to the level of risk of AI. |
| Strengthening international cooperation | Strengthen the system for international cooperation on the development of international norms for data protection in the field of AI. |

International

G7

G7 DATA PROTECTION AND PRIVACY AUTHORITIES' STATEMENT ON GENERATIVE AI (JUNE 2023)

On 21 June 2023, data protection authorities from the Group of Seven (G7) countries met for a roundtable in Japan on developments and challenges from generative AI technologies from the perspective of data protection and privacy. Following the roundtable, the authorities jointly released a “**Statement on Generative AI**,”⁴⁰ outlining the substantive areas of agreement from their discussion.

Notably, the Statement recognizes that existing laws apply to generative AI products and highlights the Italian data protection authority's enforcement action against OpenAI.

The Statement highlights several issues under existing data protection and privacy laws that may arise in the context of generative AI. These include:

- » **Legal authority for the processing of personal data**, particularly that of minors and children, in relation to:
 - the datasets used to train, validate and test generative AI models;
 - individuals' interactions with generative AI tools; and
 - the content generated by generative AI tools.
- » **Security safeguards** to protect against threats and attacks that seek to:
 - invert the generative AI model to extract or reproduce personal information originally processed in the datasets used to train the model; and
 - subvert the efficacy of measures designed to promote compliance with other privacy and data protection requirements.

» **Mitigation and monitoring measures** to ensure personal information generated by generative AI tools is:

- **accurate, complete and up-to-date**; and
- free from **discriminatory, unlawful, or otherwise unjustifiable effects**.

» **Transparency measures** to promote openness and explainability in the operation of generative AI tools, especially in cases where such tools are used to make or assist in decision-making about individuals.

» **Production of technical documentation** across the development lifecycle to **assess the compliance** of generative AI tools with privacy and data protection requirements.

» Technical and organizational measures to ensure **individuals** affected by or interacting with these systems have the **ability to exercise their rights** in relation to generative AI tools with respect to:

- access to their personal information;
- rectification of inaccurate personal information;
- erasure of their personal information; and
- refusal to be subject to solely automated decisions with significant effects.

» **Accountability measures** to ensure appropriate levels of responsibility among actors in the AI supply chain, especially when generative AI models are built upon one another.

» **Limiting collection of personal data** to only that which is necessary to fulfill the specified task.

The Statement also **recommends practices** that developers and providers of generative AI systems should employ to embed the concept of “Privacy by Design” in the design, conception, operation, and management of new products and services that use generative AI technologies.

These recommendations include:

- » Complying with existing laws.
- » Adhering to applicable internationally observed **data protection and privacy principles**, such as:
 - data minimization;
 - data quality;
 - purpose specification;
 - use limitation;
 - security safeguards;
 - transparency;
 - rights for data subjects, including the right to be informed about the collection and the use of their personal data, and
 - accountability.
- » Documenting conception, operation, and management choices and analyses in a **privacy impact assessment**.
- » Putting in place **measures to ensure that deployers or adopters of generative AI systems are also able to comply** with their data protection and privacy obligations.

HIROSHIMA AI PROCESS COMPREHENSIVE POLICY FRAMEWORK (DECEMBER 2023)

On 1 December 2023, digital and technology ministers from the G7 countries, together with the OECD and the Global Partnership on AI, endorsed the “Hiroshima AI Process Comprehensive Policy Framework.”⁴¹

According to the ministers’ statement, the Framework is the culmination of work within the Hiroshima AI Process under Japan’s G7 Presidency and includes the following:

- » the OECD’s Report towards a G7 Common Understanding on Generative AI;⁴²
- » the “**International Guiding Principles for Organizations Developing Advanced AI Systems**” (International Guiding Principles)⁴³
- » the voluntary “**International Code of Conduct for Organizations Developing Advanced AI Systems**” (International Code of Conduct)⁴⁴ and
- » project-based cooperation on AI.

The **International Guiding Principles** are intended to contribute to the development of a comprehensive policy framework for advanced AI systems.

The **International Code of Conduct** outlines actions that organizations are encouraged to take to give effect to International Guiding Principles, in line with a risk-based approach.

Both documents aim to promote safe, secure, and trustworthy AI worldwide. Building on the existing OECD AI Principles, they are designed to provide non-exhaustive guidance to “**organizations**” developing “**advanced AI systems**.”

- » “**Organizations**” may include, among others, entities from academia, civil society, the private sector, and the public sector.
- » “**Advanced AI systems**” are defined to include the most advanced foundation models and generative AI systems.

They apply to all AI actors, when and as relevant, during the design, development, deployment and use of advanced AI systems.

| International Guiding Principles | | Code of Conduct |
|--|--|---|
| Principle | Elaboration of Principle | |
| 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 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. | <p>Testing should take place in a secure environment 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.</p> <p>The Code of Conduct highlights several risks that should be considered in designing and implementing testing measures:</p> <ul style="list-style-type: none"> » Chemical, biological, radiological, and nuclear risks. » Offensive cyber capabilities. » Risks to health and/or safety. » Risks from models of making copies of themselves or “self-replicating” or training other models. » Societal risks, including harmful bias and discrimination or infringement of legal frameworks, including data protection. » Threats to democratic values and human rights, including the facilitation of disinformation or harming privacy. » Risks 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. <p>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.</p> <p>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.</p> <p>These measures should be documented and supported by regularly updated technical documentation.</p> |

| International Guiding Principles | | Code of Conduct |
|--|--|--|
| Principle | Elaboration of Principle | |
| 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. 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. | Bounty systems, contests, or prizes could be used to incentivize the responsible disclosure of weaknesses. |
| Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increased accountability. | 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 as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately; also, transparency reporting should be supported and informed by robust documentation processes. | <p>This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems.</p> <p>These reports, instruction for use, and relevant technical documentation, as appropriate, should be kept up-to-date and should include, for example;</p> <ul style="list-style-type: none"> » 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. <p>Robust documentation processes include technical documentation and instructions for use.</p> |

| International Guiding Principles | | Code of Conduct |
|--|---|---|
| Principle | Elaboration of Principle | |
| 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. | <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> |
| 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, where feasible throughout the AI lifecycle. | <p>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.</p> <p>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.</p> <p>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.</p> <p>Organizations should establish policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices.</p> |

| International Guiding Principles | | Code of Conduct |
|---|---|--|
| Principle | Elaboration of Principle | |
| Invest in and implement robust security controls, including physical security, cybersecurity, and insider threat safeguards across the AI lifecycle. | <p>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.</p> | <p>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.</p> <p>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.</p> |
| 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. | <p>This includes, where appropriate and technically feasible, content authentication such as 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.</p> <p>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.</p> | <p>Organizations should collaborate and invest in research, as appropriate, to advance the state of the field.</p> |

| International Guiding Principles | | Code of Conduct |
|--|---|---|
| Principle | Elaboration of Principle | |
| 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 of 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. |
| 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. | 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. |
| Advance the development of and, where appropriate, adoption of 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 Standards Development Organizations (SDOs). | 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. |
| 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 transparency of training datasets should also be supported, and organizations should comply with applicable legal frameworks. | 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. |

US Executive Order on the Safe, Secure, and Trustworthy Development of AI (October 2023)

Author: Lee Matheson

This section benefited from review and recommendations by Amie Stepanovich.

On October 30, 2023, U.S. President Joe Biden signed Executive Order 14110, “**Executive Order on the Safe, Secure, and Trustworthy Development of Artificial Intelligence**” (“EO 14110” or the “EO”).⁴⁵

EO 14110 defines the current administration’s policy on AI, following two earlier Executive Orders signed by the previous administration. Under U.S. law, an Executive Order is a lawfully binding directive issued by the President of the United States to the executive agencies under the President’s capacity to manage agencies’ staff and resources, and constitutional authority to execute the laws of the United States. Executive Orders remain in force until they are canceled or superseded by a future Executive Order, adjudicated unlawful by court, or expire based on their own terms. There is no automatic expiration process for such orders.

According to an accompanying Fact Sheet from the White House,⁴⁶ EO 14110 “establishes new standards for AI safety and security, protects Americans’ privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.” As might be expected, an effort to make such comprehensive rules runs nearly 60 pages. EO 14110 also follows a previous publication from the Biden White House, the Office of Science and Technology Policy’s (OSTP) 2022 Blueprint for an AI Bill of Rights. It is also important to note that, immediately after the publication of EO 14110, the Office of Management and Budget (OMB) published a draft policy on government agency use of AI, which was finalized in March 2024.⁴⁷

Key Definitions under the EO

“Artificial Intelligence” is “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.”

A “dual use foundation model” is “an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range

of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters.”

AI and the U.S. Federal Government

The Executive Order primarily governs the procurement, development, and use of AI and policies related to AI within and by the federal government, calling for U.S. federal agencies to engage in the creation of both generally applicable government-wide safety standards and agency-specific AI safety requirements. Agencies are generally directed to follow eight guiding principles when undertaking the EO’s directives and are also required to undertake a few specific actions – for example, all agencies are required to designate a Chief AI Officer within 60 days of the EO’s publication, and the Director of the OMB is ordered to provide a list of recommendations that will be required from vendors seeking to fulfill AI contracts. The Chief AI Officer of each agency will be responsible for creating internal AI governance bodies, developing agencies’ compliance plans, and creating AI use case inventories.⁴⁸

The EO also mandates the Department of Labor to assess the impact of AI on the labor market, and to develop, publish, and adopt principles and best practices to mitigate potential AI-driven harms such as worker displacement and employers’ AI-related collection and use of employee data.

In addition to the above, the EO contains provisions that govern government procurement of personal information from data brokers. The EO approaches this issue from a privacy perspective, mandating that the Director of OMB “evaluate and take steps to identify” the acquisition of “commercially available information” (CAI) by agencies – particularly when such data contains personal information – and create “appropriate agency inventory and reporting processes.” Ultimately, OMB is directed to work with other government agencies to create guidance to agencies on how to mitigate privacy and confidentiality risks stemming from agency use of CAI.

Implications for Industry

The EO has significant implications, both directly and indirectly, for industry as well as government agencies. For one, it directs the U.S. National Institute of Standards and Technology (NIST) to lead an effort that will establish guidelines and best

practices “with the aim of promoting consensus” on AI safety throughout industry. The EO also directs the Secretary of Commerce to create systems, **including private sector reporting requirements**, to monitor the safety of “certain large AI models” and to solicit public input on potential risks, benefits, and policy approaches for “certain foundation models.” The Secretary of Commerce is ultimately directed to draft a report to inform the President of their findings.

The Secretary of Commerce is additionally directed to both determine a set of conditions defining when a large AI model might be used maliciously, and to develop know-your-customer (KYC) requirements that will apply to specific providers of Infrastructure as a Service (IaaS) products and require them to report when their products are used by foreign persons to train large AI models that have “potential” to be used in malicious activity. These directives may raise privacy and data protection concerns, as they may require collection of significant information about the customers of AI providers.

Other testing and transparency obligations include:

- » Requiring vendor companies to meet transparency requirements and disclose to the government prior to use.
- » Requiring companies developing “dual-use foundation models” to provide “safety reports” to the government, including the results of the red-team testing mandated by the new NIST guidance.
- » Requiring companies to report the acquisition or development of “large-scale computing clusters” to the government – the exact threshold triggering reporting left to a future collaborative effort between the Secretary of State, Secretary of Defense, Secretary of Energy, and the Director of National Intelligence.

The EO is also likely to impact private industry indirectly in many ways. One provision of the EO calls upon the U.S. Office of Management and Budget (OMB) to “develop an initial means to ensure that agency contracts for the acquisition of AI systems and services” align with other principles of the EO. In 2024, OMB published a request for information on Responsible Procurement for AI in Government, initiating a process that could define practices, standards, and contractual requirements for government acquisition of AI technologies.⁴⁹ Government procurement implicates hundreds of billions of dollars each year, and standards and guidelines developed for the procurement of AI systems by US government entities will implicate any organization wishing to pursue the U.S. government as a client.⁵⁰ Further, even entities who do not wish to imminently pursue procurement contracts may decide to implement the same standards such that they may qualify for a future contract opportunity.

Civil Rights and Equity

Another significant theme in the EO is its recognition of the implications of AI for civil rights issues, including in areas of criminal justice, access to government benefits and programs, and in the broader economy. Regarding criminal justice, the EO directs the Attorney General of the United States to coordinate a “cohesive effort” across government agencies to address algorithmic discrimination and produce a set of best practices to mitigate when AI is used in the criminal justice system. Concerning government benefits, the civil rights offices of each agency are directed to identify how AI is being used to administer benefits and address any unlawful discrimination that is resulting from that use. Several agencies with particularly sensitive mandates, including the Consumer Finance Protection Bureau and the Department of Housing and Urban Development, are directed to take particular risk mitigation steps to address the particular impact of AI within the industries that they regulate.

International Cooperation

The EO does not pretend that the United States is creating AI policy in a void. Several government bodies are ordered to monitor and influence the use of AI by foreign governments and other global actors; the Secretary of State is also ordered to articulate the United States’ position on the role of AI in global development.

What’s Next?

Because of the EO’s focus on “consequential impacts” and “significant effects” and its reliance on developing internal agency AI expertise, it is likely that the coming months and years will see a significant number of guidance documents published by the White House as well as other government agencies. Notably, the White House maintains a record of its recent activities related to AI, and several federal agencies have already published or announced AI guidance relevant to their respective areas of responsibility. These documents all share analyses of the implications of existing federal laws for various uses of AI as well as forward-looking priorities.⁵¹

Executive Order 14110 is not the only executive action the United States has taken related to Artificial Intelligence. On February 28, 2024, the White House issued Executive Order 14117, specifically to prevent bulk access to U.S. persons’ sensitive data by strategic adversaries of the United States, specifically citing the potential use of such data in the development of AI as one of the concerns addressed by the EO.⁵²

European Union Artificial Intelligence Act

Author: Bianca-Ioana Marcu

This section benefited from review and recommendations by Vasileios Rovilos.

In April 2021, the European Commission unveiled the proposal for a “Regulation Laying Down Harmonised Rules on Artificial Intelligence” (AI Act), recognizing the potential of AI systems to bring societal and economic growth, as well as the need to regulate the potential harms arising from their deployment. On 13 March 2024, the AI Act was formally approved by the European Parliament⁵³ and is expected to enter into force in June 2024 as the world’s first horizontal, binding regulation on AI.

The AI Act forms a core part of an existing framework of European laws regulating the digital environment, including rules governing the processing of personal data and the provision of digital services to European citizens. Within this context, the AI Act will introduce a set of obligations for both developers and deployers of AI to ensure:

- » A **well-functioning internal market for AI** in the European Union (EU);
- » That AI systems are **safe, trustworthy**, and **respect fundamental rights and values**.

The AI Act will apply in all EU Member States and could have broad extraterritorial application to non-EU entities developing and deploying AI systems for the EU market.

The AI Act defines AI as “a machine-based system that is designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.” While there is no universal definition of AI, there are efforts to align regional and international definitions in order to create a coherent regulatory environment for the deployment of AI technologies.


A risk-based approach to regulating AI

One of the cornerstones of the AI Act is its risk-based approach, founded on a classification system determining the level of risk that the technology could pose to a person’s health, safety, and fundamental rights. On this basis, the AI Act proposes a set of scalable rules which vary from banning certain applications of AI to providing heightened obligations for AI applications deemed to be high-risk, to requiring voluntary codes of conduct.

The AI Act defines five levels of risk in AI:

- » **Unacceptable risk** – AI systems that are considered to pose a clear threat to the health, safety and rights of people will be banned. Examples of prohibited practices include social scoring by governments, real-time biometric identification systems in public spaces, and biometric categorization systems using a person’s sensitive characteristics.
- » **High-risk** – AI systems that are considered to pose a high risk to the health, safety, and rights of people, and will be subject to strict obligations before they can be placed on the market. Examples of high-risk practices include AI-powered critical infrastructure systems, the use of AI in the provision of essential public and private services, and the administration of justice.
- » **Systemic risk** – The notion of systemic risk is applicable in the context of a general-purpose AI (generative AI) model if it has “high impact capabilities” or if it is based on a decision of the European Commission. In the context of general-purpose AI models, the concept of “high impact capabilities” can be determined on the basis of appropriate technical tools and methodologies, including indicators and benchmarking.
- » **Limited risk** – AI systems which pose a limited risk to the health, safety, and rights of people will have to be accompanied by specific transparency obligations. Examples of limited risk applications include AI-enabled chatbots.
- » **Minimal or no risk** – The proposal allows for the free use of AI systems which pose minimal or no risk, for example AI-enabled videogames and spam filters. In this instance, the proposal encourages the adoption of voluntary codes of conduct.

With its risk-based approach, the AI Act will introduce the obligation for providers of high-risk AI systems to conduct a Conformity Assessment (CA). The CA is a legal obligation that must be performed prior to placing an AI system on the EU market. The CA is designed to foster accountability and transparency, and to identify and mitigate risks posed by high-risk AI. Conducting a CA includes several requirements that providers of high-risk AI must embed in the design of such systems throughout their lifecycle, including maintaining a **risk management** system, ensuring **high quality of data sets**, maintaining **technical documentation**, and ensuring sufficient **transparency**. Furthermore, high-risk AI systems must have an appropriate level of **accuracy, robustness**, and **cybersecurity**.



Once the CA requirements are duly completed and implemented, the provider of the high-risk AI system draws up an EU declaration of conformity and affixes the CE marking. The CA process can be done either internally (by the provider) or externally, by a ‘third-party’ (notified bodies).

The AI Act and Generative AI

Providers of general-purpose AI systems, including generative AI models, will have to comply with a specific set of rules under the AI Act, including EU copyright law. Providers of general-purpose AI systems will have to draw up and maintain technical documentation of the model, with details regarding the training and testing process of the system and the results of its evaluation. Moreover, providers will, among other obligations, have to make publicly available a detailed summary about the content used to train the general-purpose AI model, and cooperate with national supervisory authorities.

The AI Act stipulates additional obligations for providers of general-purpose AI systems *with systemic risk*, one of the levels of risk described above. These additional obligations include performing model evaluation (including conducting adversarial testing), assessing and mitigating possible systemic risk at the Union level, reporting serious incidents and corrective measures taken to address them, and ensuring an adequate level of cybersecurity.

How and when will the AI Act be enforced?

Providers of high-risk AI systems, complying with the CA process specified above, will be supervised by the notified bodies (as designated by the notifying authorities) of the EU Member States. Furthermore, the AI Act will establish a “European Artificial Intelligence Board” (European AI Board) that will be tasked with ensuring effective cooperation between national supervisory authorities and the European Commission, issue guidance and analyses on the AI Act, and assist in ensuring the consistent application of the law.

Enforcement of the obligations vested on providers of general-purpose AI models *with systemic risk* will be a task for the European Commission’s AI Office. The AI Office will monitor, supervise, and enforce the

AI Act requirements on general-purpose AI models and systems across EU Member States. To support the implementation and enforcement of the AI Act, a scientific panel of independent experts will be established.

As the AI Act is set to enter into force in June 2024, important milestones towards the implementation of the law include:

- » **6 months after its entry into force** – The general provisions (pertaining to scope and definitions) will become applicable. Moreover, the provisions on prohibited AI practices will also be applicable.
- » **12 months after its entry into force** – The provisions on (newly launched) general-purpose AI will be applicable. However, general-purpose AI models pre-dating the AI Act will have 3 years to comply with said provisions. Additionally, within this time frame, EU Member States will have to appoint their market surveillance authorities.
- » **No later than 18 months after its entry into force** – The European Commission (after consulting the European AI Board) has to provide guidelines specifying the practical implementation for the classification of high-risk AI systems.
- » **24 months after its entry into force** – Mark the general applicability of the provisions of the AI Act.
- » **36 months after its entry into force** – The obligations for high-risk AI systems, as included in Annex I, will become applicable. Additionally, (pre-existing) general-purpose AI models will have to comply with the set provisions.
- » **72 months after its entry into force** – The obligations set for high-risk AI will also be applicable to pre-existing high-risk AI systems used by public authorities.

Furthermore, the designed penalties for non-compliance with the AI Act are significant, particularly for non-compliance with the provisions on prohibited AI practices (administrative fines of up to €35 million) and on high-risk AI systems (administrative fines of up to €15 million).

APPENDIX ENDNOTES

- 1 <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework>
- 2 https://www.chiefscientist.gov.au/sites/default/files/2023-06/Rapid%20Response%20Information%20Report%20-%20Generative%20AI%20v1_1.pdf
- 3 <https://consult.industry.gov.au/supporting-responsible-ai>
- 4 https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/Safe-and-responsible-AI-in-Australia-discussion-paper.pdf
- 5 https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf
- 6 <https://www.esafety.gov.au/newsroom/media-releases/new-industry-recommendations-to-curb-harms-of-generative-ai>
- 7 <https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai>
- 8 <https://www.esafety.gov.au/industry/tech-trends-and-challenges>
- 9 <https://dp-reg.gov.au>
- 10 <https://www.oaic.gov.au/newsroom/digital-platform-regulators-release-working-papers-on-algorithms-and-ai>
- 11 <https://dp-reg.gov.au/publications/working-paper-2-examination-technology-large-language-models>
- 12 https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html. An unofficial English translation is available at <https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/>
- 13 https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. An unofficial English translation is available at <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>
- 14 https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html. An unofficial English translation is available at <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/>
- 15 http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm
- 16 http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
- 17 <https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf>
- 18 https://www.gov.cn/gongbao/content/2020/content_5492511.htm
- 19 https://www.gov.cn/zhengce/content/202306/content_6884925.htm
- 20 <http://www.fxcw.org.cn/dyna/content.php?id=26910>
- 21 <https://www8.cao.go.jp/cstp/ai/ningen/ningen.html>. English translation available at <https://www8.cao.go.jp/cstp/ai/humancentricai.pdf>
- 22 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf
- 23 https://www.ppc.go.jp/files/pdf/230602_kouhou_houdou.pdf
- 24 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240119_report.html
- 25 <https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html>
- 26 <https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework>
- 27 https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf
- 28 <https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/singapore-launches-ai-verify-foundation-to-shape-the-future-of-international-ai-standards-through-collaboration>
- 29 <https://aiverifyfoundation.sg/what-is-ai-verify/>
- 30 <https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox>
- 31 https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf
- 32 https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf
- 33 Available at https://openresearch-repository.anu.edu.au/bitstream/1885/277585/1/SKAI_31.pdf. No authoritative English language translation is available.
- 34 <https://eiec.kdi.re.kr/policy/materialView.do?num=208784&topic=&pp=20&datecount=&recommend=&pg=>
- 35 <https://www.koreaherald.com/view.php?ud=20201223000794>
- 36 https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_Y2B1M0R6G2I2P1B0V2X9H4Z0X3M3J2
- 37 <https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&pageIndex=7&bbsSeqNo=94&nttSeqNo=3183387&searchOpt=ALL&searchTxt=>
- 38 <https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9055#LINK>
- 39 <https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9083>
- 40 https://www.priv.gc.ca/en/opc-news/speeches/2023/s-d_20230621_g7/
- 41 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document02_en.pdf
- 42 <https://www.oecd.org/publications/g7-hiroshima-process-on-generative-artificial-intelligence-ai-bf3c0c60-en.htm>
- 43 <https://www.mofa.go.jp/files/100573471.pdf>
- 44 <https://www.mofa.go.jp/files/100573473.pdf>
- 45 Full text available at: https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf?utm_campaign=subscription+mailing-list&utm_medium=email&utm_source=federalregister.gov
- 46 White House Fact Sheet, “President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence,” available at <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>.
- 47 Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, WHITE HOUSE OFFICE OF MANAGEMENT AND BUDGET, available at <https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf>; see also White House Fact Sheet, “Vice President Harris Announces OMB Policy To Advance Governance, Innovation, and Risk Management in Federal Agencies Use of Artificial Intelligence,” available at <https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/28/fact-sheet-vice-president-harris-announces-omb-policy-to-advance-governance-innovation-and-risk-management-in-federal-agencies-use-of-artificial-intelligence/>.
- 48 A comprehensive list of the obligations imposed on agencies by the EO may be found here: <https://crsreports.congress.gov/product/pdf/R/R47843>
- 49 Request for Information: Responsible Procurement of Artificial Intelligence in Government, WHITE HOUSE OFFICE OF MANAGEMENT AND BUDGET, 89 FR 22196, available at <https://www.federalregister.gov/documents/2024/03/29/2024-06547/request-for-information-responsible-procurement-of-artificial-intelligence-in-government>
- 50 A Snapshot of Government-Wide Contracting for 2022, U.S. GOVERNMENT ACCOUNTABILITY OFFICE, available at <https://www.gao.gov/blog/snapshot-government-wide-contracting-fy-2022>
- 51 See Administration Actions on AI, available at <https://ai.gov/actions/>; see also FTC Proposes New Protections to Combat AI Impersonation of Individuals, available at <https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals>; Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964, U.S. EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, available at <https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial>; *Civil Rights in the Digital Age: The Intersection of Artificial Intelligence, Employment Decisions, and Protection Civil Rights*, U.S. Department of Justice, JOURNAL OF FEDERAL LAW AND PRACTICE, Vol. 70, no. 1, pp. 57-68, available at <https://www.justice.gov/file/1189116/d?inline=>
- 52 Executive Order 14117, Preventing Access to Americans’ Bulk Sensitive Personal Data and United States Related Data by Countries of Concern, Executive Office of the President, available at <https://www.federalregister.gov/documents/2024/03/01/2024-04573/preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-related-data-by-countries-of-concern>
- 53 <https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law>

