Navigating Governance Frameworks for Generative AI Systems in the Asia-Pacific

MAY 2024

FUTURE OF PRIVACY FORUM | AI

AUTHORED BY

Dominic Paulger
Policy Manager for Asia-Pacific, Future of Privacy Forum

ACKNOWLEDGEMENTS

This paper benefited from review and recommendations by Gabriela Zanfir-Fortuna, Anne J. Flanagan, Rob van Eijk, and Josh Lee Kok Thong. It also benefited from review and contributions from Bianca-Ioana Marcu and Lee Matheson.

FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

About Future of Privacy Forum (FPF)

The **Future of Privacy Forum (FPF)** is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

About FPF's Center for Artificial Intelligence

The **Center for Artificial Intelligence** at the Future of Privacy Forum is dedicated to navigating the complex landscape of AI governance and its intersection with privacy and data protection law. Drawing on expertise from a global Leadership Council comprising industry leaders, academics, civil society, and policymakers, the Center provides sophisticated, practical policy analysis to help organizations align innovation with responsible implementation while meeting evolving regulatory requirements. Learn more about the FPF Center for AI at fpf.org/ai.

TABLE OF CONTENTS

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Across the APAC region, there is increasing interest in both understanding how generative artificial intelligence (AI) systems and large language models (LLMs) work, and exploring approaches to manage these technologies.

Leveraging the Future of Privacy Forum (FPF)'s global work on governance and regulation of AI,[1] FPF's Asia-Pacific (APAC) office commenced a research project on the regulatory and governance landscape for generative AI systems and LLMs in the APAC region in April 2023. The project focuses on 5 jurisdictions:

1. **Australia**
2. **China**
3. **Japan**
4. **Singapore**
5. **South Korea**

This Report is the culmination of that project. It notes that these jurisdictions are at an inflection point in the governance of generative AI systems, with a risk of fragmentation both within and between jurisdictions as regulatory responses diverge.

Section 1 of the Report charts early regulatory responses to generative AI in the 5 APAC jurisdictions. It notes that while most responses to date have favored voluntary guidelines and multi-stakeholder consultations, China has taken a unique approach by enacting binding regulations for generative AI.

This section also posits that as jurisdictions develop their respective generative AI governance frameworks, there are areas of consensus that could inform efforts to address regulatory fragmentation. Specifically, the 5 jurisdictions broadly agree on five identifiable risks posed by generative AI systems, and on certain recommended courses of action to address these risks. These five risks are:

- factual inaccuracies;
- lack of transparency;
- inappropriate use of personal data;
- malicious use of generative AI systems; and
- biased or discriminatory output.

Section 2 examines the broader landscape of existing laws and regulations in the five jurisdictions that may apply to generative AI. It highlights data protection law as a key source of legal obligations for developers and deployers of generative AI systems due to the use of personal data in training these systems.

This section also discusses data protection issues that are relevant for generative AI, such as lawful grounds for collecting and processing personal data, including publicly available personal data, managing data quality, handling personal data breaches, and fulfilling individual rights such as access to, correction, and erasure of personal data.

Section 3 highlights takeaways for policymakers and developers and deployers of generative AI, including but not limited to those in industry, in the APAC region to foster responsible governance of generative AI.

A key takeaway for **policymakers** is the need to counter the risk of regulatory fragmentation across the region. This may include ensuring alignment and interoperability with international policy frameworks, providing guidance on applying existing laws to generative AI, and promoting cross-regulator coordination.

Amongst the key takeaways for **developers and deployers of generative AI** are that robust internal AI governance structures, data management practices, privacy protection processes, security safeguards, and transparency measures are widely recognized building blocks for responsible development and deployment of generative AI systems. Enabling traceability of AI-generated content and clearly indicating its nature are also unanimous recommendations across the early regulatory responses to generative AI in the five jurisdictions.

INTRODUCTION

This Report explores the regulatory and governance landscape for generative AI in the APAC region, focusing on 5 jurisdictions: Australia, China, Japan, Singapore, and South Korea.

This Report observes that policymakers in the APAC region have generally taken a different approach to AI governance to their counterparts in other regions. Whereas several jurisdictions, such as the European Union (EU), have worked on crafting AI specific laws,[2] the APAC region has generally prioritized voluntary frameworks and non-binding guidelines.

Although generative AI has been around in some form for decades, it was only in late 2022 that publicly accessible and consumer-facing generative AI systems entered the market at scale. Policymakers in APAC and around the world are still at a very early stage in developing governance frameworks that are specific to the recent generative AI boom.[3] **Section 1** of this Report charts **early regulatory responses to generative AI** in the five APAC jurisdictions (these responses are summarized in detail in the **Appendix**) and concludes that:

- The majority of policymakers in the five jurisdictions do not currently appear to be looking to enact binding regulations to govern generative AI. A key exception is China, which has enacted technology-specific regulations. South Korea also plans to enact comprehensive AI regulation, but an existing bill does not specifically address generative AI.
- There are areas of consensus in emerging regulatory responses to generative AI among the five jurisdictions in APAC. These areas of consensus could inform efforts by policymakers and industry to increase regulatory interoperability between jurisdictions, as APAC jurisdictions develop their governance approaches for generative AI. Examples of such areas of consensus include identified common risks and recommended measures in the emerging generative AI regulatory frameworks.

Importantly, the development of generative AI systems does not take place in a regulatory vacuum. In the absence of binding regulations that specifically address generative AI, the main sources of legal obligations for generative AI are relevant existing technology-neutral laws and regulations. To that end, **Section 2** of the Report charts existing laws and regulations in the 5 jurisdictions that may apply to generative AI today. Findings include:

- **Data protection law has been a key source of binding legal obligations for generative AI.** This is because datasets containing personal data have been used to train several existing generative AI models, and this has therefore brought the resulting systems within the scope of such law.
- There are **several data protection issues particularly relevant for generative AI** in the 5 jurisdictions, such as the lawful grounds for processing personal data, or the rules applicable to collecting and processing publicly available personal data.
- In APAC (as elsewhere), **data protection authorities have been uniquely placed to take regulatory action** regarding generative AI systems, and some have already taken such action or issued guidance.

Looking to the future, **Section 3** of the Report notes that the APAC region is at an inflection point in the governance of generative AI systems. In particular, the section highlights that there is a risk of regulatory fragmentation surrounding generative AI in APAC because:

- *within* jurisdictions, multiple laws, frameworks, and guidelines may apply to generative AI systems; and
- *between* jurisdictions, regulatory responses to generative AI diverge as each jurisdiction pursues different priorities.

Finally, the Report highlights takeaways for policymakers and developers and deployers of generative AI, to consider in developing governance frameworks for generative AI and addressing the risk of regulatory fragmentation within and between jurisdictions.

Notes

This Report is informed by discussions held during two roundtables organized by FPF APAC in 2023, which sought input on the project from regulators, industry leaders, academics, and civil society representatives from across the APAC region, and beyond.

- The first roundtable was jointly organized with the Personal Data Protection Commission of Singapore (PDPC) and held in person during Singapore's Personal Data Protection Week in July 2023.
- The second roundtable was held virtually in October 2023 to open the discussion to a broader range of stakeholders from across the APAC region.

Both roundtables were held under the Chatham House Rule. FPF sincerely thanks all stakeholders who participated in our roundtables for their participation and insights.

The Report that follows should not be taken to reflect the views of any participant in the roundtables, and any errors are attributable to the author. The Report does not constitute legal advice.

Scope

The Report focuses on responsibilities for private sector organizations that develop and deploy AI under general laws, frameworks, and guidelines that apply to all kinds of organizations. It does not focus on sectoral AI laws or frameworks (e.g., healthcare, financial services, or other similar highly regulated sectors) or consider the responsibilities of public sector bodies.

Further, it also does not focus on: (1) intellectual property issues; (2) environmental, social, and governance risks; (3) competition concerns; and (4) labor force issues.

Definitions

This Report uses the term "**generative AI system**" expansively to include systems built using generative AI models, as well as applications built on top of such models.

This Report uses the term "**governance**" to refer to the set of policies and procedures that seek to ensure that AI technologies are developed, deployed, and used responsibly. This includes both voluntary frameworks and legally binding obligations.

GENERATIVE AI

Generative AI is a subset of AI technology that involves **AI models** – programs that have been trained on a set of data to recognize patterns or make decisions without further human intervention.[4] The models that power generative AI systems are capable of producing content in response to open-ended instructions from users in the form of a "prompt." The responses can appear human-made and are different with each iteration – the same prompt can produce a different output each time it is given to the system, especially given the iterative nature of models that continue to be trained by the data contained in the prompts.

These factors set generative AI apart from earlier "discriminative" AI models[5] that are effective at identifying the distinctions between different classes of data and are well-suited for tasks like pattern recognition, data classification, and prediction.

Although generative AI systems often appear to have human-like creative abilities, it is important to note that in producing content, generative AI models are merely analyzing patterns in their training data and using this analysis to make predictions – for instance, about what word to place next in a sentence or what pixel to place in an image. The models that power generative AI systems were not designed to store and retrieve information with 100% accuracy or verify the accuracy of the information they produce.[6] However, generative AI systems can combine these models with other technological solutions that allow for limited verification.

Modern generative AI systems can accept inputs and produce outputs in a wide range of different "modes," based on the data that they have been trained on.

LLMs are traditionally trained on large amounts of text data and so can accept inputs and produce outputs in text. More recently, several developers have begun releasing "**multimodal**" models, such as OpenAI's GPT-4, Google's Gemini, and Anthropic's Claude, which can accept inputs and produce outputs in both text and image form.

An ongoing area of research focuses on broadening the scope of generative AI models to incorporate additional data modalities, such as 3D environments, video, and audio. For instance, in February 2024, OpenAI introduced its Sora model, which enables users to generate a minute-long video from a short text prompt.[7]

When it comes to considering the policy implications of generative AI, it is helpful to look at how the technology is typically built. A key point that stakeholders raised during FPF's two roundtables on the governance of generative AI systems was the need to consider how generative AI is built across the different layers of its technology stack. We elaborate more on the technology stack for generative AI below.

Infrastructure layer

Generative AI systems run on physical hardware, demanding large capacity server infrastructure systems incorporating incredibly powerful central processing units (CPUs) and graphics processing units (GPUs). As most mainstream generative AI systems require such significant amounts of computational power, these systems often run high-density, environmentally controlled servers in cloud data centers which may be centrally located or distributed across different physical locations.

Model layer

Modern generative AI is built on **new AI models** that are based on complex algorithms that are highly capable and have been found to demonstrate human-like performance on a wide range of tasks. The technological development that enabled the development of these models was the discovery of a new type of neural network algorithm, known as "**transformers**" in 2017.[8] Combined with advances in computing infrastructure, the transformer algorithm allowed for the creation of AI **models** that can be trained efficiently on massive amounts of data.

Further, while much attention has been paid to so-called "general purpose"[9] generative AI models (such as OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude), there are narrower, use-case specific models that are trained on highly specific forms of data:

- **computer code** (e.g., Codex (OpenAI);[10] AlphaCode (DeepMind);[11] Project Wisdom (IBM)).[12]
- **chemistry data** (e.g., ChemBERTa (University of Toronto);[13] Chemformer (AstraZeneca);[14] MoLFormer (IBM));[15]
- **climate data** (e.g., ClimaX (Microsoft);[16] IBM-NASA geospatial intelligence foundation model);[17] and
- **financial data** (e.g., BloombergGPT (Bloomberg));[18]
- **ancient texts**;[19]
- **protein structures**;[20] and
- **organic molecules.**[21]

Application layer

The majority of users do not interact directly with generative AI models, but rather with **applications** that are built upon these models.

Applications that are built on generative AI models make highly capable AI accessible to the public at large and have the potential to aid humans in conducting numerous tasks like computer coding, content creation, transcription, and translation that used to be laborious or required significant training.

A tipping point for the adoption of applications built on generative AI models was the public release of an LLM-based chatbot, ChatGPT, by OpenAI in November 2022.[22] It was perhaps the first time in history that the public had interacted with an AI application that was both available for widespread consumer use and able to respond to or perform a wide range of different tasks, rather than just specific tasks.[23] Since then, an increasing number of such applications have been released on the market for a wide range of different use cases.

SECTION 1

Regulatory Responses to Generative AI in APAC

Overview

Unlike their counterparts in the EU who have worked on crafting AI-specific laws and hard regulations, such as the **AI Act**, policymakers in the APAC region have generally taken a different approach to AI governance, prioritizing voluntary frameworks and non-binding guidelines.

APAC Frameworks for AI Generally

Prior to 2023, AI-specific governance frameworks in the 5 jurisdictions covered by this Report were mainly limited to voluntary ethical principles and guidelines that apply generally to all forms of AI, including but not limited to generative AI. However, they were drafted before the emergence of modern generative AI systems so do not take into account specific policy considerations that are unique to such technologies. These include:

- **Australia's "AI Ethics Framework" (2019)**;[24]
- **China's "Ethical Principles for New Generation AI" (2021)**;[25]
- **Japan's "Social Principles of Human-Centric AI," (2019)**,[26] **and "Governance Guidelines for Implementation of AI Principles" (2022)**;[27]
- **Singapore's "Model AI Governance Framework" (2020)**;[28] and
- **South Korea's "Human Centered AI Ethics Standards" (2020)**.[29]

Of the 5 jurisdictions, only **South Korea** has been actively working on a comprehensive AI regulation. A draft "Bill on Fostering AI and Creating a Foundation of Trust" was tabled in the National Assembly in June 2021, but has not been enacted as of this Report's publication.[30]

Further information on the content of these frameworks may be found in the Appendix.

APAC Frameworks for Generative AI

Following the public release of several major generative AI systems in early 2023, the 5 jurisdictions have taken a wide range of different regulatory responses to generative AI, reflecting their distinct priorities, legal frameworks, and the role they envision for these technologies.

These responses are outlined below. Further information on the approaches adopted by each jurisdiction may be found in the Appendix.

AUSTRALIA

- **"Rapid Response Information Report on Generative AI" (March 2023)**, an influential report commissioned by the Australian Government that provides an overview of the development, regulatory landscape, and potential risks and opportunities of LLMs and foundation models.[31]
- **The eSafety Commissioner's "Tech Trends Position Statement on Generative AI" (August 2023)**, which provides an explainer on generative AI technologies and guidance for industry on minimizing online harm risks when developing and deploying generative AI.[32]
- **A public consultation on "Safe and Responsible AI in Australia"** that included a discussion paper published in June 2023,[33] and an interim response from the Australian government in January 2024.[34]
- **"Digital Platform Regulators Forum Working Paper on LLMs" (October 2023)**, a paper examining the regulatory implications of LLMs across several regulatory domains, including privacy and online safety.[35]

CHINA

- **"Regulations on the Administration of Deep Synthesis of Internet Information Technology"** (**Deep Synthesis Regulations**) **(January 2023)**, a set of binding regulations applying to deep-fakes and other forms of synthetic media.[36]
- **"Interim Measures for the Management of Generative AI Services" (Interim Generative AI Measures) (August 2023)**, a more detailed regulation outlining state policy principles for generative AI and establishing obligations on service providers throughout the lifecycle of a generative AI system.[37]
- **"Basic Security Requirements for Generative AI Services" (February 2024)**, a technical standard that outlines technical requirements for complying with the Interim Generative AI Measures.[38]

JAPAN

- **"Notice Regarding Cautionary Measures on the Use of Generative AI Services" (June 2023)**, a short guideline on complying with Japanese data protection law when using LLM chatbots, issued by the Personal Information Protection Commission (PPC) together with an **enforcement decision against OpenAI**.[39]
- **"Guidelines for AI Business Operators" (April 2024)**, a set of draft guidelines that aim to update Japan's voluntary framework in response to advanced AI, including generative AI.[40]

SINGAPORE

- **"Generative AI: Implications for Trust and Governance" (June 2023)**, a discussion paper that outlines proposals for policymakers and business leaders to build a trusted, responsible global ecosystem for adopting generative AI.[41]
- **"Proposed Model AI Governance Framework for Generative AI" (January 2024)**, a draft policy framework that builds on the earlier discussion paper that highlights potential regulatory actions and governance measures various stakeholders could adopt to enhance generative AI trust and safety.[42]

SOUTH KOREA

- The Personal Information Protection Commission (PIPC)'s **enforcement decision against Open AI (March 2023)**.[43]
- **"Policy Direction for Safe Use of Personal Information in the AI Era" (August 2023)**, which outlines measures that the PIPC will take in response to emerging privacy challenges from AI and provides preliminary guidance on protecting privacy when developing and deploying AI systems.[44]

Comparison Shows a Wide Spectrum of Policy Responses

This subsection of the Report (1) compares the policy responses to generative AI in the 5 jurisdictions; summarizes (2) risks from generative AI identified by these policy responses; and (3) measures proposed by these policy responses to govern generative AI.

The majority of the above policy responses to generative AI are voluntary governance frameworks or early efforts by policymakers to shape the future direction of governance of generative AI. To date, only China has enacted legally binding regulations. That said, these policy responses mark an evolution away from earlier, principle-based or thematic frameworks to more comprehensive frameworks that increasingly recognize that different responsibilities may arise at the different stages of the AI lifecycle.

Many of these policy responses also identify specific **risks** arising from generative AI and, in some cases, propose **measures** that developers and deployers could adopt to improve their governance of generative AI.

These are covered in more detail below in the subsections on risks identified and measures recommended by policymakers in the 5 jurisdictions.

While the diversity of these responses precludes a fully like-for-like comparison, a high-level comparison of these approaches reveals a spectrum of policy strategies, from voluntary guidelines and international collaboration, to the development of comprehensive national legislation and actions by various regulators.

Jurisdictions Differ in the Forms and Legal Effect of Their Policy Responses to Generative AI

A key difference between jurisdictions is in whether they prioritize voluntary norms and bottom-up multi-stakeholder frameworks or binding top-down regulations for governing generative AI.

Broadly, Australia, Singapore, and Japan have favored **multi-stakeholder consultations** involving government, industry, academia, and civil society. Their aim is to develop **internationally aligned frameworks** and **voluntary guidelines** that enable responsible innovation in generative AI while mitigating risks:

- Australia's approach has prioritized domestic public consultation, leveraging expert reports, and inter-agency coordination to establish a risk-based framework.
- Japan has based its approach on international collaboration, notably through its G7 presidency in 2023.
- Singapore has sought to bring together local, regional, and international stakeholders to collaborate on developing a bespoke governance framework for generative AI, as well as on AI governance testing for generative AI systems (and other AI technologies more broadly).

In contrast, China has adopted a more prescriptive approach by enacting two sets of **binding, technologically specific regulations** to govern

generative AI and related issues, such as synthetic media. These aim to align the provision of generative AI-powered services to China's national interests and principles and impose obligations on providers of services using generative AI and technologies that use AI to create or modify media.

South Korea is charting a hybrid course: while its Ministry of Science and ICT has been developing comprehensive national AI legislation, its data protection authority, the PIPC, has concurrently focused on issuing detailed guidance and establishing programs to enable AI innovation while managing privacy risks.

Jurisdictions Differ in How Their Policy Responses to Generative AI Allocate Roles and Responsibilities

Another key difference concerns the allocation of roles and responsibilities within emerging governance frameworks. Policy responses to generative AI in the 5 jurisdictions are increasingly recognizing that different responsibilities arise at different stages of the lifecycle of an AI system, including development and deployment.

However, a comparison of emerging governance frameworks across the 5 jurisdictions reveals that although the frameworks recognize these different responsibilities, they do not always clearly identify the different roles associated with them.

For instance, some frameworks clearly differentiate between the responsibilities of developers, deployers, and users at different stages of the AI lifecycle. Frameworks in this category include Japan's Guidelines for AI Business Operators and to a lesser extent, Singapore's Proposed Model AI Governance Framework for Generative AI.

Others, however, identify responsibilities that apply at different stages of the AI lifecycle but use the generic term "service providers" for all such responsibilities, without differentiating between those who could be categorized elsewhere as developers or deployers. Frameworks in this category include the Tech Trends Position Statement on Generative AI by Australia's eSafety Commissioner and China's generative AI regulations.

Jurisdictions Differ as to Which Entities Have Issued Responses to Generative AI

Within jurisdictions, a further difference is in which agencies or branches of government have been leading efforts to govern generative AI.

AUSTRALIA

The Australian government has taken a multi-agency approach. The **Department of Industry, Science and Resources** (DISR) leads public consultations and framework development.

Australia's data protection authority, the **Office of the Australian Information Commissioner** (OAIC)'s role in governance of generative AI has largely been confined to its participation in the **Digital Platform Regulators Forum** (DP-REG), and it has not issued any generative AI-specific guidance to date.

By contrast, another DP-REG member, the **eSafety Commissioner,** has played a far more active role by providing guidance to industry on mitigating risks from generative AI, albeit with a focus on online safety.

CHINA

China's approach is centralized. China's cyberspace regulator, the **Cyberspace Administration of China** (CAC) has been responsible for issuing all binding regulations. Technical standards are set by the cybersecurity standards body, known as **TC260**.

JAPAN

Japan has prioritized international collaboration through the G7 Hiroshima AI Process. These efforts have involved the **Ministry of Foreign Affairs**, the **Ministry of Internal Affairs and Communications**, the **Digital Agency**, and the **Ministry of Economy, Trade and Industry** (METI).

Domestically, **METI** has played a central role in developing AI governance frameworks, while the data protection authority, the **PPC**, has issued preliminary guidance on the privacy implications of the use of LLM chatbots.

SINGAPORE

Singapore's combined infocommunications, media, and data protection regulator, the **Infocomm Media Development Authority** (IMDA) has led collaborative efforts, partnering with industry to establish the AI Verify Foundation and engaging stakeholders through initiatives like the proposed generative AI governance framework.

SOUTH KOREA

The **Ministry of Science and ICT** (MSIT) has been leading the development of a comprehensive AI bill.

The data protection authority, the **PIPC**, has taken a proactive stance on data protection and enforcement actions against non-compliant AI practices.

Data protection authorities are often in a unique position relative to other agencies or branches of government as they have an existing legal mandate to regulate the processing of personal data. This

enables them to regulate generative AI systems that are trained on or otherwise use personal data. In APAC, data protection authorities in Japan and South Korea have been most active in addressing generative AI's privacy and data protection implications. Both jurisdictions issued guidance for businesses on complying with data protection laws when using services like ChatGPT. Both have also pursued enforcement actions against OpenAI over ChatGPT's handling of personal data.

Common Risks from Generative AI Identified by Policymakers in the 5 Jurisdictions

As policymakers in the 5 jurisdictions covered by this Report seek to understand and respond to this evolving landscape, they have highlighted several potential risks from generative AI in their various regulatory responses.

Given the rapid pace of advancement in generative AI capabilities, it is fair to assume that not all potential risks posed by generative AI have been identified at this stage. However, an analysis of APAC policymakers' responses to generative AI indicates an emerging consensus around key risks. In particular, policymakers in all 5 of the jurisdictions covered by this Report identified the following risks from existing generative AI systems:

- the potential for generative AI systems to produce **factual inaccuracies**;[45]
- **lack of trust and transparency**;
- **inappropriate use of personal data** to train generative AI models;
- **malicious use** of generative AI systems, including to spread misinformation and disinformation; and
- generation of **biased or discriminatory content**.

Factual Inaccuracies

Policymakers highlighted risks associated with generative AI's tendency to produce factual inaccuracies. Several regulatory responses raise concerns that such inaccuracies can mislead or even harm people – for instance, by making defamatory statements that harm individuals' reputations or providing ineffective health advice.

This risk arises because generative AI models are probabilistic rather than deterministic – they generate output by predicting what item should come next in a sequence (e.g., a word in a sentence or a pixel in an image) based on the data that they have been trained on. Text-based generative AI applications may therefore sometimes produce statements that are grammatically correct but factually inaccurate based on probabilities assigned to information in their training data.

While the models can make highly accurate predictions, they have no grounding in the real world outside of their training data. This means that, by default, these systems are unable to verify the information they produce and may fail to understand the context of that information. For instance, the information may not have been fact-checked, may reflect deliberate misinformation posted online, or may not have been intended to be factually accurate, as is the case for creative works like fiction, poetry, or humor.[46] Even when trained on accurate statements, LLMs can still occasionally assign high probabilities to factually inaccurate statements due to their reliance on statistical patterns.[47]

Lack of Trust and Transparency

Policymakers highlighted lack of transparency as a risk for generative AI systems. Transparency is viewed as a multifaceted issue spanning:

- documenting model and system capabilities, training data, limitations, and intended uses;
- explaining organizational policies; and
- clearly indicating when people are interacting with AI systems or AI-generated content.

Crucially, policymakers in all 5 of the jurisdictions have closely linked transparency with accountability and have highlighted that without insight into how these systems work and what data they were trained on, it becomes difficult to understand outputs, apportion liability, seek redress, anticipate safety risks, and establish effective safeguards.

Potential transparency gaps identified include lack of clarity around personal data use, incomplete transparency on model capabilities and limitations, and inadequate information for stakeholders to make informed decisions and establish effective safeguards.

Inappropriate Use of Personal Data

Policymakers highlighted that the training of generative AI models on personal data may give rise to data protection and privacy risks.

Notably, policymakers raised particular concerns where such data was obtained through "**scraping**" – the automated extraction of data from the internet. For instance, in South Korea, the PIPC's Policy

Direction for Safe Use of Personal Information in the AI Era highlights that generative AI systems that have been trained on "scraped" personal data may effectively process personal data in ways unanticipated by data subjects i.e. to train Generative AI models, thereby potentially increasing the scale of privacy infringements.

This aligns with broader international trends. For instance, in August 2023, data protection authorities from 12 jurisdictions, including Australia, released a **joint statement on data scraping and the protection of privacy**.[48] The statement outlines privacy risks related to data scraping, including targeted cyberattacks, identity fraud, monitoring, profiling, unauthorized political or intelligence gathering, and unwanted direct marketing. The statement also emphasizes that even publicly available personal data remains subject to data protection laws, with obligations applying to both data scrapers and operators of platforms hosting the data.

Other risks arise from the tendency of generative AI models to "**memorize**" specific phrases, sentences, or even longer passages from their training data and reproduce this information in their outputs.[49] This can lead systems inadvertently to leak personal data, as well as confidential information, that they have been trained on.[50] This data security risk may also extend to personal data that users input into a generative AI system, such as an LLM chatbot, if the system retains that data for further training.

This risk is discussed in further detail under Section 2: Existing Laws.

Malicious Use

Policymakers recognized the risk of users and threat actors attempting to circumvent safeguards designed to prevent the malicious use of generative AI systems. This practice, commonly known as "**jailbreaking**," involves manipulating prompts to bypass restrictions and make AI perform undesired tasks.[51] Research has shown that it can be relatively easy to evade safeguards established by generative AI service providers by slightly altering the prompt. For instance, a generative AI system might be prohibited from explaining how to rob a bank *per se* but may still provide the prohibited information if asked to write a one-act play about how to rob a bank, or to explain how to rob a bank "for educational purposes."[52]

Policymakers have highlighted the following as potential malicious uses for the technology, including:

- **Fraud**: AI-generated content can be used to facilitate "phishing" attacks, where threat actors deceive individuals into disclosing sensitive information by posing as trustworthy entities.
- **Promotion of hate, discrimination, and abuse**: Generative AI can be manipulated to spread harmful ideologies and promote abusive behavior.
- **Harmful content**: In the absence of proper guardrails, AI systems may inadvertently provide inappropriate or harmful information, potentially causing harm to users or facilitating unlawful acts.
- **Abusive and illegal content**: Recent advancements in image generation models have enabled the creation and dissemination of abusive and illegal content, such as the non-consensual creation of synthetic media featuring the likenesses of real persons in compromising situations, and the creation of synthetic child sexual abuse material.
- **Malware creation**: Generative AI models have the potential to be used in crafting malicious software code across various programming languages for cybercrime purposes.
- **Misinformation and disinformation**: Most policymakers highlighted the risk that generative AI could be used maliciously to increase the scale and effectiveness of misinformation and disinformation campaigns.

Bias and Discrimination

Statistically, it is likely that any data set will contain some bias. Policymakers have therefore identified the risk that biases in the data used to train generative AI systems may cause these systems to produce outputs that amplify these biases and encourage discrimination.

Broadly, policymakers highlight two high-risk forms of bias that can arise in generative AI training data. **Historical bias** refers to patterns of harmful stereotypes and negative attitudes towards certain groups being reflected in the training data due to historical depictions of those groups.[53] **Representation bias** occurs when certain groups are over or underrepresented in the datasets.[54] Both types of bias in a training data set can potentially result in discriminatory output which can be harmful to individuals. In the case of generative AI, where the algorithm is constantly being trained by its human reviewers, this harm may be further amplified.

A related issue is "**toxicity**."[55] Where the training dataset contains negative, discriminatory, offensive, or excessively insensitive perspectives, a generative AI system may produce outputs that reflect such perspectives, even without direct human influence. When it comes to determining what forms of content are toxic, such assessment depends on the context and is often highly subjective. Additionally, some forms of content may be offensive even if they do not use blatantly inflammatory language, as is the case with coded language or "dog whistles."

Measures Recommended by Policymakers in the 5 Jurisdictions to Govern Generative AI Vary in Nature, but Share Some Commonalities

In addition to identifying potential risks from generative AI, policymakers in all 5 jurisdictions covered by this Report have begun identifying possible measures that developers and deployers of generative AI could implement to address these potential risks.

However, policymakers have done so in a range of different contexts, for different purposes, and with different levels of granularity.

Policymakers in Australia (DISR) and Singapore have focused on **proposing, and seeking feedback on, potential measures** that could be included in future AI guidelines or regulation. These proposals tend to be drafted in very broad terms, and businesses are not necessarily expected to adopt them.

Further, policymakers in Australia (eSafety Commissioner), Japan, and South Korea have focused on issuing **voluntary guidelines recommending specific measures** for industry to implement. These guidelines target different groups, including:

- all businesses that develop, deploy, and use AI (Japan);
- those businesses who process personal data in the development, deployment, and use of generative AI (South Korea);
- those whose development, deployment, and use of generative AI has implications for users' online safety (Australia).

These guidelines outline recommended measures in detail but are still sufficiently open-ended and flexible that different businesses could adapt them to their specific needs.

Policymakers in China have focused on **enacting legally binding regulations requiring service providers to implement measures** in relation to specific technologies.

The diversity of these responses precludes a fully like-for-like comparison. However, a broad survey of regulatory responses to generative AI in the 5 jurisdictions covered in this Report indicates that there are some emerging areas of substantial consensus.

In particular, all 5 of the jurisdictions highlighted the need for organizations to develop their own **internal AI governance and risk management policies** and **provide transparency** by publishing documentation (e.g., model cards), AI governance policies, and transparency reports.

The survey also indicates further areas of consensus in measures highlighted by at least 4 of the 5 jurisdictions. These include:

- Conducting **impact assessments** to identify and mitigate harm.
- Implementing measures to **manage data quality** to mitigate against harmful biases.
- Developing **privacy management programs** and disclosing a privacy policy.
- Deploying **security measures**, including:
  - Assessing security risks;
  - Conducting security testing on systems before deployment;
  - Monitoring systems after deployment;
  - implementing measures to address identified risks and vulnerabilities;
  - Reporting security incidents;
  - Implementing security controls to address risks to both physical security and cybersecurity;
  - Sharing information regarding risks and best practices.
- Developing and implementing **measures to indicate that content is AI-generated**, such as watermarking, labeling, and other authentication and provenance mechanisms.

Interestingly, many of these measures overlap significantly with recommendations made by the G7 in the "Hiroshima AI Process Comprehensive Policy Framework" – one of the most detailed international frameworks for governance of advanced AI systems, including generative AI, that has been released at the international level to date. As policymakers in the APAC region continue to develop their national-level generative AI governance frameworks, there is scope for international alignment along these lines.

SECTION 2

Existing Laws in the 5 Jurisdictions Likely Relevant to Generative AI

The majority of the jurisdictions covered by this Report currently lack AI-specific laws, such as the EU's AI Act and their regulatory responses to generative AI generally have not been legally enforceable.

This means that for the time being, the main source of binding legal obligations governing generative AI systems in all jurisdictions covered by this Report (except China) remains existing, technology-neutral laws. Further, even for jurisdictions like China that have begun to enact regulations to address specific risks presented by generative AI, existing laws will remain relevant to matters that fall outside of the scope of these generative AI-specific regulations.

Identifying all relevant laws and obligations across the 5 jurisdictions is beyond the scope of this Report and would depend on the specific circumstances of how organizations develop, deploy, and use generative AI.

The table below provides a **snapshot of existing legal frameworks** in the 5 jurisdictions covered by this Report that may be relevant to risks from generative AI identified by policymakers and summarized in the previous section, from consumer protection law to criminal law.

Importantly, the mapping in the table below does not include data protection and privacy law, whose cross-cutting applicability to generative AI will be explored in detail in the following sub-section.

Mapping of Existing Legal Frameworks in the 5 Jurisdictions that are Relevant to Generative AI, in addition to Data Protection Law

POTENTIAL HARM CAUSED BY GENERATIVE AI SYSTEMS	AUSTRALIA	CHINA	JAPAN	SINGAPORE	SOUTH KOREA
Producing factually inaccurate or misleading output	**Consumer protection law** Competition and Consumer Act: Australian Consumer Law **Civil remedies** Common law of contract and tort (e.g., negligence, defamation, misrepresentation) **Professional regulation** (finance, medical, legal sectors)	**Consumer protection law** Law on the Protection of Consumer Rights and Interests **Civil law** (contract, tort, defamation, etc.) Civil Code **Professional regulation** (finance, medical, legal sectors)	**Consumer protection law** Act against Unjustifiable Premiums and Misleading Representations **Civil law** (contract, tort, defamation, etc.) Civil Code **Professional regulation** (finance, medical, legal sectors)	**Consumer protection law** Consumer Protection (Fair Trading) Act **Civil remedies** Common law of contract and tort (e.g., negligence, defamation, misrepresentation) Defamation Act **Professional regulation** (finance, medical, legal sectors)	**Consumer protection law** Framework Act on Consumers Act on Fair Labeling and Advertising Act on the Regulation of Terms and Conditions **Civil law** (contract, tort, defamation, etc.) Civil Act **Professional regulation** (finance, medical, legal sectors)
Misplaced human reliance on AI-generated content	**Civil remedies** Common law of contract and tort (e.g., negligence) **Professional regulation** (finance, medical, legal sectors)	**Civil law** (contract, tort, etc.) Civil Code **Professional regulation** (finance, medical, legal sectors)	**Civil law** (contract, tort, etc.) Civil Code **Professional regulation** (finance, medical, legal sectors)	**Civil remedies** Common law of contract and tort (e.g., negligence) **Professional regulation** (finance, medical, legal sectors)	**Civil law** (contract, tort, etc.) Civil Act **Professional regulation** (finance, medical, legal sectors)

POTENTIAL HARM CAUSED BY GENERATIVE AI SYSTEMS	AUSTRALIA	CHINA	JAPAN	SINGAPORE	SOUTH KOREA
Causing physical economic, or psychological harm	**Criminal law** Criminal Code **Consumer protection law** Competition and Consumer Act, Australian Consumer Law **Civil remedies** Common law of contract and tort. **Sectoral laws** (finance, medical, legal sectors) **Online Safety** Online Safety Act	**Criminal law** Criminal Law **Consumer protection law** Law on the Protection of Consumer Rights and Interests **Civil law** (contract, tort, etc.) Civil Code Tort Liability Law **Sectoral laws** (finance, medical, legal sectors)	**Criminal law** Penal Code **Consumer protection law** Consumer Product Safety Act **Civil law** (contract, tort, etc.) Civil Code **Sectoral laws** (finance, medical, legal sectors)	**Criminal law** Penal Code **Consumer protection law** Consumer Protection (Fair Trading) Act **Civil remedies** Common law of contract and tort **Sectoral laws** (finance, medical, legal sectors) **Online safety law** Online Safety Code Online Criminal Harms Act 2023	**Criminal law** Criminal Act **Consumer protection law** Framework Act on Consumers Act on the Regulation of Terms and Conditions **Civil law** (contract, tort, etc.) Civil Act **Sectoral laws** (finance, medical, legal sectors) **Online content law** Telecommunications Business Act Network Act
Creating biased or discriminatory content	**Civil remedies** Common law of contract and tort. **Anti-discrimination law** Racial Discrimination Act Sex Discrimination Act Disability Discrimination Act Age Discrimination Act	**Criminal law** Criminal Law **Civil law** (contract, tort, etc.) Civil Code **Content regulation** Regulations on Ecological Governance of Internet Information Content	**Criminal law** Penal Code **Civil law** (contract, tort, etc.) Civil Code **Anti-Discrimination law** Act on the Promotion of Efforts to Eliminate Unfair Discriminatory Speech and Behavior Against Persons Originating from Outside Japan	**Criminal law** Penal Code **Civil remedies** Common law of contract and tort. Protection from Harassment Act 2014. **Anti-Discrimination law** Maintenance of Religious Harmony Act	**Criminal law** Criminal Act **Civil law** (contract, tort, etc.) Civil Act
Creating/ spreading disinformation or misinformation	**Criminal law** Criminal Code **Laws against the spread of online disinformation and misinformation** Australian Code of Practice on Disinformation and Misinformation (voluntary)[56] *Combatting Misinformation and Disinformation Bill* **Civil remedies** Common law of tort (e.g., defamation)	**Criminal law** Criminal Law, Articles 221, 243 **Laws against the spread of online disinformation and misinformation** Regulations on Ecological Governance of Internet Information Content **Civil law** (contract, tort - defamation, etc.)	**Criminal law** Penal Code **Civil law** (contract, tort - defamation, etc.) Civil Code	**Criminal law** Penal Code Defamation Act **Laws against the spread of online disinformation and misinformation** Protection from Online Falsehoods and Manipulation Act 2019 **Civil remedies** Common law of tort (e.g., defamation). Defamation Act	**Criminal law** Criminal Act **Online content law** Telecommunications Business Act Network Act **Civil law** (contract, tort - defamation, etc.) Civil Act

POTENTIAL HARM CAUSED BY GENERATIVE AI SYSTEMS	AUSTRALIA	CHINA	JAPAN	SINGAPORE	SOUTH KOREA
Bullying and harassment	**Criminal law** Criminal Code, Sections 474.17 and 474.17A (using a carriage service to menace, harass or cause offense) **Civil remedies** Common law of tort **Online Safety** Online Safety Act, Parts 5 (cyber-bullying material targeted at an Australian child), 6 (non-consensual sharing of intimate images), 7 (cyber-abuse material targeted at an Australian adult)	**Civil law** (e.g. tort) Civil Code **Anti-harassment regulation** Law on the Protection of Women's Rights and Interests *Draft law on cyberbullying*	**Criminal law** Penal Code **Civil law** (e.g. tort) Civil Code **Anti-harassment regulation** Equal Opportunity Act Act on the Promotion of Efforts to Eliminate Unfair Discriminatory Speech and Behavior Against Persons Originating from Outside Japan	**Criminal law** Penal Code **Civil remedies** Common law of tort **Online safety law** Online Safety Code **Anti-harassment regulation** Protection from Harassment Act 2014	**Criminal law** Criminal Act **Civil law** (e.g. tort) Civil Act **Online content law** Telecommunications Business Act Network Act
Fraud, including phishing	**Criminal law** Division 134 (Obtaining property or a financial advantage by deception) **Civil remedies** Common law of tort. **Online Safety** Online Safety Act	**Criminal law** Criminal Law, Articles 266, 287 Law on Combating Telecom and Online Fraud **Civil law** (e.g. tort) Civil Code	**Criminal law** Act on the Prohibition of Unauthorized Computer Access, Articles 6, 7, 12 Penal Code, Article 161-2 **Civil law** (e.g. tort) Civil Code	**Criminal law** Penal Code, ss 416, 419, 170 Computer Misuse Act 1993, ss 3-4 **Civil remedies** Common law of tort. **Online safety law** Online Criminal Harms Act 2023	**Criminal law** Criminal Code **Civil law** (e.g. tort) Civil Act **Online content law** Telecommunications Business Act Network Act
Malware generation	**Criminal law** Criminal Code, Vol 2, Part 10.7 (computer offenses)	**Criminal law** Criminal Law, Articles 285-287 Cybersecurity Law, Article 27	**Criminal law** Penal Code, Article 161-2, 168-2, 168-3, 234-2, 246-2	**Criminal law** Computer Misuse Act 1993, s 5 Computer Misuse and Cybersecurity (Amendment) Act 2017, s 8B(1)(b)	**Criminal law** Criminal Act

POTENTIAL HARM CAUSED BY GENERATIVE AI SYSTEMS	AUSTRALIA	CHINA	JAPAN	SINGAPORE	SOUTH KOREA
Creation/ distribution of abusive material » Extremist content » Child abuse » Non-consensual intimate images	**Criminal law** Criminal Code, Divisions 80 (urging violence and advocating terrorism or genocide), 471, D (Offences relating to use of carriage service for child abuse material), 474 (telecommunications offences) **Civil remedies** Common law of contract and tort **Data protection law** Privacy Act **Online Safety** Online Safety Act, Parts 6 (non-consensual sharing of intimate images), 8 (material that depicts abhorrent violent conduct), 9 (online content scheme)	**Civil law (e.g. tort)** Civil Code **Data protection law** Personal Information Protection Law **Online safety law** Regulations on Ecological Governance of Internet Information Content	**Criminal law** Penal Code Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children **Civil law** (e.g. tort) Civil Code **Data protection law** Act on the Protection of Personal Information	**Criminal law** **Civil remedies** Common law of contract and tort. **Data protection law** Personal Data Protection Act **Online safety law** Online Safety Code Online Criminal Harms Act 2023	**Criminal law** Criminal Act **Civil law** (e.g. tort) Civil Act **Data protection law** Personal Information Protection Act **Online content law** Telecommunications Business Act Network Act
Age-Inappropriate Content	**Online Safety** Online Safety Act, Parts 9 (online content scheme)	**Online Safety** Law on the Protection of Minors Regulations on Ecological Governance of Internet Information Content	**Online Safety** The Act on Development of an Environment that Providers Safe and Secure Internet Use for Young People	**Online Safety law** Online Safety Code Online Criminal Harms Act 2023	**Online content law** Telecommunications Business Act Network Act Youth Protection Act

Data Protection

Though numerous existing laws and regulation in the 5 jurisdictions may apply to generative AI, data protection law is likely to be one of the main sources of binding legal obligations for generative AI, considering the horizontal applicability of the rules to any "processing of personal data," the fact that there are dedicated supervisory authorities to enforce it, and the common use of personal data in training generative AI models.

While only some of the 5 jurisdictions reviewed in this Report have laws specifically designed to address areas like online safety or misinformation, all have data protection laws. These include:

- **Australia**'s Privacy Act, which gives effect to the Australian Privacy Principles (APPs).[57]
- **China**'s Personal Information Protection Law (PIPL).[58]
- **Japan**'s Act on the Protection of Personal Information (APPI).[59]
- **Singapore**'s Personal Data Protection Act (PDPA).[60]
- **South Korea**'s Personal Information Protection Act (PIPA).[61]

Further, these laws have generally already been implemented. Regulatory authorities have been established to enforce data protection laws in all of the 5 jurisdictions, and many have also issued detailed guidance on compliance.

This subsection of the Report raises several considerations that:

- policymakers can think through when examining how their respective data protection laws apply to generative AI, and
- developers and deployers of generative AI systems can note when complying with data protection laws in the 5 jurisdictions.

Not every instance of operating a generative AI system will involve the processing of personal data. However, many generative AI systems do so, especially where the large datasets used to train the generative AI model contain personal data, and/or the system collects personal data that users have entered into it, and uses this data to further train the underlying model.

Insofar as generative AI models or applications process personal data, they may be subject to obligations under personal data protection law. Failure to comply with these obligations may give rise to penalties or other sanctions from data protection authorities.

There have already been several important decisions by data protection authorities in Italy, Japan, and South Korea concerning OpenAI's provision of services to users through its LLM chatbot, ChatGPT.[62]

Common issues across these decisions included:

- Lack of legal authority to process personal data to train a generative AI model.
- Failure to adequately inform data subjects about the processing of their personal data, including failing to provide information in languages other than English.
- Processing personal data in violation of data protection principles, such as data quality and data minimization.
- Failing to provide mechanisms for data subjects to exercise rights, such as correction of their data or opting out of processing of their data to train the GPT model.
- Failing to verify the ages of users and obtain parental consent for use of ChatGPT by minors.

Further, data protection authorities in **Japan** and **South Korea** have issued guidance that specifically addresses the application of data protection laws to generative AI**.**

Separately, there have also been statements from multiple data protection authorities at the international level that identify issues under existing data protection and privacy laws that may arise from the development and deployment of generative AI. These statements include:

- **the G7 data protection and privacy authorities' statement on generative AI (June 2023)**
- **the Global Privacy Assembly's Resolution on Generative AI (October 2023); and**
- **the joint data protection authorities' (DPAs) statement on data scraping (August 2023)**.

Legal Authority to Process Personal Data to Train Generative AI Models

One of the biggest challenges for organizations to comply with existing data protection laws in the context of developing and deploying generative AI is ensuring that these organizations have legal authority (or 'lawful ground,' or 'legal basis') to process personal data to train generative AI models.

All data protection laws require organizations to fulfill certain criteria (such as obtaining **consent** from data subjects, or establishing that the processing is **necessary** for a specified purpose) before organizations have legal authority to process personal data.

Previous work by FPF has identified two main challenges for organizations that process personal data in multiple jurisdictions in APAC.[63] These include: (1) the lack of consistency between jurisdictions in the

available legal bases for processing personal data; and (2) the lack of alternative legal bases to consent (such as "legitimate interests") that can be relied on to process personal data in a variety of different circumstances.

This work found that as a result of these issues, organizations that process personal data in multiple APAC jurisdictions would likely have to build their compliance frameworks around consent, as this legal basis was the main "common denominator" for data protection laws in the APAC region, especially for sensitive personal data.

The most suitable legal basis for processing personal data will likely depend on the circumstances in which the organization obtained the dataset for training a generative AI model. These circumstances include:

- **Collecting data** through "**web crawls**" of publicly available web pages containing personal data (also commonly referred to as "scraping");
- **Collecting data from end-users of generative AI applications to refine the underlying model**. Generative AI applications may also process personal data if end-users input personal data via a prompt, and the application retains that data (e.g., to further train the underlying AI model); and
- **Reusing an existing dataset** that contains personal data.

COLLECTION OF PERSONAL DATA

Training Datasets Obtained through "Web Crawls"

Modern generative AI systems, particularly LLMs, rely heavily on training data derived from large-scale "crawls" of the internet.[64] This involves employing automated programs to systematically navigate and extract information from websites.

For instance, several common, publicly available datasets for training AI systems are based on the "**Common Crawl**" – a massive compilation of publicly available websites collected regularly since 2008.[65] These include:

- **Colossal Clean Crawled Corpus** (C4), a cleaned version of the Common Crawl prepared by AllenAI that has been used to train, among others, Meta's LLaMa model.[66]
- **LAION-400M**[67] **and LAION-5B**:[68] Two datasets containing, respectively, 400 million and 5 billion pairs of image and text data prepared by the **Large-scale Artificial Intelligence Open Network** (LAION) that was used to train, among others, Stability AI's Stable Diffusion model.

These massive datasets are then fed into the LLM, where it learns the underlying statistical relationships and patterns within human language. This allows the AI to develop the ability to generate human-like text, translate languages, and perform other complex natural language processing tasks.

As web crawls may, by their nature, capture personal data that has been posted online, organizations would likely need to establish legal authority to process that personal data by fulfilling the requirements for a legal basis under relevant data protection laws.

However, there are practical issues with doing so. Processed data may have been made public without those individuals' knowledge or consent; it may also qualify as sensitive under various data protection laws, particularly if such data has been leaked online. Further, given the size of crawled datasets, such data may potentially relate to millions of data subjects worldwide.

Given these factors, it may not be possible to rely on **consent** to process personal data in training datasets obtained through web crawls. First, it would not be feasible to identify the data subjects whose personal data is present in the dataset. Second, even if an organization was able to identify such data subjects, the organization may not have the necessary contact information to seek their consent as it has no prior relationship with them.

More suitable legal bases within the 5 jurisdictions studied are those which permit the processing of personal data without consent if:

- the personal data is **publicly available**;
- the processing is necessary for a **legitimate interest** of the organization or a third party; or
- the processing is for **research purposes**.

Consent and Sensitive Personal Data: As discussed above, it is likely infeasible to obtain consent from data subjects in this situation. Another challenge is that data protection laws in 4 of the 5 jurisdictions covered by this Report (Australia, China, Japan, and South Korea) require consent to process **sensitive personal data**.

3 of these 4 jurisdictions (Australia, China, and South Korea) do not provide alternatives to consent that would apply to the use of personal data to train a generative AI model. This could prevent organizations operating in these jurisdictions from using web crawled datasets for this purpose or otherwise expose them to the risk of enforcement actions from data protection authorities.

Publicly Available Personal Data: Data protection laws in 3 of the 5 jurisdictions (China, Japan, and Singapore) expressly provide legal bases that permit the processing of publicly available personal data without consent. Organizations would likely have little difficulty in complying with the relevant provisions in Japan's APPI and Singapore's PDPA to process personal data for the purpose of training a generative

AI model, as these provisions appear to only require that the personal data is available to the public.

However, organizations may encounter difficulties relying on the relevant provision in China's PIPL due to the safeguards required. In particular, it may be difficult to argue that the processing of a data subject's publicly available personal data for the purpose of training a generative AI model would not have a significant impact on the data subject's rights and interests, given that internationally, there have been several high-profile cases where generative AI systems trained on publicly available data have produced factual inaccuracies that are potentially defamatory. For instance, in April 2023,

- The Washington Post reported that ChatGPT had falsely claimed that a law professor had been accused of sexual harassment and cited a non-existent Washington Post article as the source of the information.[69] The newspaper came out to say that there was no such article.
- An Australian mayor threatened to sue OpenAI for defamation after ChatGPT falsely claimed that he had been convicted of bribery and imprisoned.[70]

Further, the relevant provision of the PIPL is of limited benefit in this situation as it does not apply to sensitive personal data.

Legitimate Interests: Data protection laws in 2 of the 5 jurisdictions (Singapore and South Korea) permit collection and use of personal data without the data subject's consent if the collection or use is in the legitimate interests of the organization. However, the relevant provision of South Korea's PIPA notably does not apply to sensitive personal data.

It would be possible for an organization to argue that the development or refinement of a generative AI system is in the legitimate interests of a developer.

However, under both laws, the organization would need to establish that this interest outweighs the rights and interests of the data subject. In the absence of guidance from data protection authorities, organizations may be reluctant to take the legal risk of relying on this interpretation. Evidence that the organization has implemented safeguards to prevent material harms to data subjects from the processing of their personal data, such as potentially defamatory AI-generated content, would likely help to bolster the organization's case that its interest outweighs the impact on the data subject.

Research Purposes: Of the data protection laws in 5 jurisdictions covered by this Report, only Japan's APPI provides an exception to consent requirements for collecting and using personal data for research purposes (the relevant provision of Singapore's PDPA applies only to use of personal data for this purpose).

Theoretically, a business could rely on this provision to process personal data to train a generative AI model provided that it collaborates with an academic institution, and one of the purposes for processing the personal data is academic research. However, these requirements may limit the value of this provision where a generative AI is trained for solely commercial purposes.

A detailed summary of relevant provisions in the data protection laws of the 5 jurisdictions is presented in the table below.

Jurisdiction	Summary of Relevant Provisions
Australia	The Privacy Act does not contain any provisions that specifically authorize the processing of personal data that is publicly available, or processing for legitimate interests or research purposes. Rather, in these situations, organizations would need to comply with the Privacy Act's requirements for collecting personal data from sources other than the data subject. In particular, the organization would have to establish that: » the data is reasonably necessary for one of its functions or activities (APP 3.1); » the collection of the personal data is by lawful and fair means (APP 3.5); » it is unreasonable or impracticable to collect personal data directly from data subjects (APP 3.6) If the personal data constitutes "**sensitive personal information**," APP 3.3 requires that the organization also obtain the data subject's **consent** for the collection of personal data. This requirement is subject to exceptions. However, none of these exceptions would generally apply to the use of personal data to train an AI model.

Jurisdiction	Summary of Relevant Provisions
China	The most relevant legal basis under the PIPL is **reasonable processing of publicly available personal data**. Article 13(6) of the PIPL permits data controllers to process publicly available personal data to a reasonable extent without the data subject's consent if: » the data subject personally disclosed the data; or » the data was otherwise legally disclosed. However, this provision is subject to two safeguards. Data controllers may not rely on this provision if: » the data subject **expressly refuses** the processing; or » processing of the publicly available data may have a **significant impact on an individual's rights and interests** (Article 27). In these cases, the data controller would have to seek **consent** from the data subject. Such consent is only valid if it is voluntarily given, explicit, and fully informed (Article 14). **Consent** is also required for the processing of **sensitive personal data** (Article 29). Additionally, the data controller must inform data subjects of why it is necessary to process such data, and what impact such processing may have on their rights and interests (Article 30).
Japan	For routine business uses of personal information, the default rule under the APPI is that businesses must **specify a legal purpose** for which they will use personal information (known as the "**purpose of use**") (Articles 17 and 19). Businesses must inform a data subject of the purpose of use either before or upon acquiring the data subject's personal information and must update the data subject if the purpose of use changes (Article 21). However, these requirements are subject to exceptions, including where informing the data subject of the purpose of use would harm the business's rights or legitimate interests or where the purpose of use is already clear in the circumstances. By default, **consent** is required for: » processing of personal information **beyond the scope necessary to achieve the purpose of use** (Article 18(2)); or » processing of "**sensitive personal information**" (Article 20(2)). However, there are **exceptions** to these requirements for **research purposes** and **publicly available data**. **Research purposes**: A business would not need to obtain consent for processing of sensitive personal information if: » the business: • obtains such information from an academic research institution; and • processes that information jointly with an academic research organization at least partially for the purposes of academic research, and » there is no risk that the processing will unjustly infringe on the data subject's rights and interests. **Publicly available data**: Additionally, the business would not need to comply with the consent requirements for processing sensitive personal information if that information is **open to the public** by a person identifiable by that information, a national government organ, a local government, an academic research institution, or other body permitted by regulations.

Jurisdiction	Summary of Relevant Provisions
Singapore	The PDPA authorizes organizations to process a data subject's personal data without consent if the processing satisfies the requirements for any of the **exceptions to consent** in the First and Second Schedules to the PDPA (Sections 13 and 17). Relevant exceptions to consent include: » Processing of personal data that is **publicly available**. » **Legitimate interests** (First Schedule, Part 3). **Publicly available data**: Part 2 of the First Schedule to the PDPA permits the collection, and use of "publicly available" personal data about an individual, without that individual's consent. Such data is considered "publicly available" if it is generally available to the public. This includes personal data which can be observed by reasonably expected means at a location or an event at which the individual appears and that is open to the public (Section 2(1)). According to guidelines from Singapore's PDPC, organizations may rely on this exception if the personal data was publicly available at the time it was collected and do not need to verify whether the data is still publicly available at the time it is used (Advisory Guidelines on Key Concepts in the PDPA, paragraph 12.87).[71] **Legitimate interests**: Part 3 of the First Schedule to the PDPA permits an organization to collect and/or use personal data if the collection and/or use is in the legitimate interests of the organization or a third party. To rely on this provision, the organization must establish that the legitimate interest outweighs any adverse effect on the individual by: » conducting a risk assessment; and » implementing reasonable measures to address any risks of adverse effects identified in the assessment.
South Korea	The most relevant legal basis under the PIPA is **legitimate interests**. Article 15(1)(6) of the PIPA permits organizations to collect and use personal data without the data subject's consent if the collection and use is necessary to achieve a legitimate interest of the organization, and that legitimate interest clearly takes precedence over the data subject's rights. Additional safeguards apply to this provision. The collection and use must be significantly related to the legitimate interest of the organization and must be within a reasonable scope. If the organization is unable to rely on this provision, it would have to obtain **consent** from the data subject. **Consent** would also be required for collection and use of **sensitive personal data** (Article 23).

Collecting Data From End-Users of Generative AI Applications to Refine the Underlying Model

Generative AI applications may collect data from prompts given to the system, which are then used to further train and refine the system. Some of these prompts may contain personal data, in which case, the application would be collecting personal data for a specific purpose and so, would be subject to the obligations of a "data controller" (or equivalent) under relevant data protection laws.

Compared with the previous scenario, where data is scraped from publicly available websites, there would be fewer issues with obtaining consent from users because the organizations would have a relationship with data subjects who use its services.

The table below presents a detailed summary of relevant legal bases for processing personal data in this scenario under the data protection laws of the 5 jurisdictions covered by this Report.

Notably, several data protection authorities' enforcement decisions against OpenAI specifically addressed this scenario (see above). These decisions all emphasized the need to obtain informed consent from users for the collection and use of personal data from prompts to further train a generative AI model and provide data subjects with the right to opt out of such collection and use.

Further, the Italian *Garante*'s preliminary order also found that OpenAI could **not** rely on a legal basis under the GDPR which allows the processing of personal data without consent, where that processing is **necessary for the performance of a contract**.

Jurisdiction	Summary of Relevant Provisions
Australia	In situations where organizations collect personal data directly from data subjects, APP 3.1 requires the organization, before collecting personal data, to establish that the data is reasonably necessary for one of its functions or activities. The collection of the personal data must also be by lawful and fair means (APP 3.5). **Consent for collection of sensitive personal data**: APP 3.3 requires that if the personal data constitutes "sensitive personal information," the organization must also obtain the data subject's **consent** for the collection of personal data. This requirement is subject to exceptions, but none of these exceptions would generally apply to use of personal data to train an AI model.
China	The most relevant legal basis under the PIPL is **consent**. In this situation, a data controller could rely on **consent** (Article 13(1). In order to be valid under the PIPL, the consent must be voluntarily given, explicit, and fully informed (Article 14). **Consent** is also required for processing of **sensitive personal data** (Article 29). Additionally, the data controller must inform data subjects of why it is necessary to process such data, and what impact such processing may have on their rights and interests (Article 30).
Japan	For routine business uses of personal information, the default rule under the APPI is that businesses must **specify a legal purpose** for which they will use personal information (known as the "**purpose of use**") (Articles 17 and 19). Businesses must inform a data subject of the purpose of use either before or on acquiring the data subject's personal information and must update the data subject if the purpose of use changes (Article 21). However, these requirements are subject to exceptions, including where informing the data subject of the purpose of use would harm the business's rights or legitimate interests or where the purpose of use is already clear in the circumstances. By default, **consent** is required for: » processing of personal information **beyond the scope necessary to achieve the purpose of use** (Article 18(2)); or » processing of "**sensitive personal information**" (Article 20(2)). PPC Japan's "Notice regarding Cautionary Measures on the Use of Generative AI Services" highlights that under Japanese data protection law, service providers should: » provide collection notices with a clear statement of the purpose(s) for which the data is collected and processed and » obtain consent from users before processing their sensitive personal information.

Jurisdiction	Summary of Relevant Provisions
Singapore	The most relevant legal bases under the PDPA are **consent** and **legitimate interests**. The PDPA authorizes organizations to collect a data subject's personal data if the organization obtains **consent** or if the collection satisfies the requirements for any of the exceptions to consent in the First and Second Schedules to the PDPA (Sections 13 and 17). **Consent**: In this situation, an organization could rely on express consent (Section 14). However, the PDPA also permits consent to be deemed under certain circumstances. A relevant circumstance to this situation is **deemed consent by notification** (Section 15A). To rely on this provision, an organization must take reasonable steps to bring to the individual's attention: » the organization's intention to process the data subject's personal data; » the purpose for which the organization will process the data; and » a reasonable period and procedure for the data subject to object to the proposed processing. The organization must also conduct an **impact assessment** to determine the likely impact of the processing on the data subject, and take steps to mitigate potential risks. **Legitimate interests**: Part 3 of the First Schedule to the PDPA permits an organization to collect and/or use personal data if the collection and/or use is in the legitimate interests of the organization or a third party. To rely on this provision, the organization must establish that the legitimate interest outweighs any adverse effect on the individual by: » conducting a risk assessment; and » implementing reasonable measures to address any risks of adverse effects identified in the assessment.
South Korea	The most relevant legal bases under the PIPA are **consent**, and **legitimate interests**. **Consent**: Article 15(1) of the PIPA permits organizations to collect and use personal data if they obtain consent from the data subject for such collection and use. Under Article 15(2) of the PIPA, when obtaining consent, the organization must inform the data subject of: » The purpose for the collection and use of the personal data. » Details of the personal data that will be collected. » The period during which the data will be retained and used. » The data subject's right to withhold consent, and the consequence of exercising the right. Under Article 15(3) of the PIPA, once the organization has obtained consent, it may use the personal data for any purpose which is within the scope reasonably related to the initial purpose for which the data was collected. **Legitimate interests**: Article 15(1)(6) of the PIPA permits organizations to collect and use personal data without the data subject's consent if the collection and use is necessary to achieve a legitimate interest of the organization, and that legitimate interest clearly takes precedence over the data subject's rights. Additional safeguards apply to this provision. The collection and use must be significantly related to the legitimate interest of the organization and must be within a reasonable scope.

REUSE OF EXISTING DATASETS

This situation assumes that an organization has legally collected personal data for a specific purpose (the **primary purpose**) but intends to use this data for a new purpose (**secondary purpose**).

In this scenario, the organization would need to ensure that it has the legal authority to use the personal data for the secondary purpose of training a generative AI model. This would depend on the legal basis relied upon to collect and use the data for the primary purpose.

It is reasonable to assume that in a business context, the legal basis to process the data for the primary purpose would likely be consent. If so, it is possible that the organization may be able to rely on this consent, if the secondary purpose is within the scope of or closely related to the primary purpose.

In other cases, the organization would have to:

» obtain **fresh consent**;

» fulfill the requirements for an alternative legal basis or exception to consent, such as **legitimate interests**; or

» anonymize the data to take it out of the scope of data protection law.

The table below presents a detailed summary of relevant legal bases for processing personal data in this scenario under the data protection laws of the 5 jurisdictions covered by this Report.

Compared with the scenario of a training dataset from a web crawl, it may be easier for the organization to obtain fresh consent, as the organization may already have established communication channels with data subjects when it sought consent to use the personal data for the primary purpose.

Jurisdiction	Summary of Relevant Provisions
Australia	The Privacy Act has specific requirements for secondary use of personal data. Specifically, APPs 6.1 and 6.2 require that if an organization holds personal data that was collected for a primary purpose, the organization may only use that data for a secondary purpose if: » the data subject **consents** to the use of their personal data for a secondary purpose, or » the secondary purpose is (directly*) **related to the primary purpose**, if the data subject would **reasonably expect** the organization to use the personal data for the secondary purpose. * For sensitive personal information.
China	Data controllers may be able to **rely on the original consent** if one of the stated purposes for processing included training of generative AI models. If not, data controllers would likely have to obtain **fresh consent** pursuant to Article 14 of the PIPL, which requires data controllers to obtain fresh consent if the purpose for processing personal information changes.
Japan	By default, Article 18(2) of the APPI requires a business to obtain the data subject's **consent** to process personal information for a secondary purpose unless such processing is within the scope necessary to achieve the primary purpose. However, this is subject to an **exception for academic research**. Businesses do not need to obtain consent to process personal data for a secondary purpose if that secondary purpose at least partially includes the purpose of academic research. To rely on this exception, the business would also have to provide the personal information to an academic research institution (or equivalent) for processing and ensure that there is no risk that the processing will unjustly infringe on the data subject's rights and interests.

Jurisdiction	Summary of Relevant Provisions
Singapore	In order to use a data subject's personal data for a secondary purpose, an organization could rely on the data subject's **consent** or a relevant exception to consent. Relevant exceptions to consent in this context may include: » **legitimate interests**; » **business improvement purposes**; and » **research purposes**. **Consent for secondary purposes**: After an organization has obtained consent to collect a data subject's personal data for a primary purpose, the organization must **notify the data subject of a secondary purpose** before using the personal data for that secondary purpose (Sections 18 and 20). If the organization collected the personal data without the data subject's consent, then the organization must either obtain fresh consent to use the personal data for a secondary purpose, or satisfy an exception to consent for use of the data. **Legitimate interests**: Part 3 of the First Schedule to the PDPA permits an organization to use personal data if the use is in the legitimate interests of the organization or a third party. To rely on this provision, the organization must establish that the legitimate interest outweighs any adverse effect on the individual by conducting a risk assessment and implementing reasonable measures to address any risks of adverse effects identified in the assessment. **Business improvement purposes**: Division 2, Part 2 of the Second Schedule to the PDPA permits an organization to use personal data for various "business improvement purposes" including improving or enhancing goods and services and developing new goods and services. To rely on this provision, the organization must establish that: » the purpose for processing cannot reasonably be achieved without the use of the personal data in an individually identifiable form; and » a reasonable person would consider the use of the personal data for that purpose to be appropriate in the circumstances. **Research purposes**: Division 3, Part 2 of the Second Schedule to the PDPA permits an organization to use personal data for a research purpose if the following conditions are met: » the research purpose cannot reasonably be accomplished unless the personal data is used in an individually identifiable form; » there is a clear public benefit to using the personal data for the research purpose; » the results of the research will not be used to make any decision that affects the individual; and » in the event that the results of the research are published, the organization publishes the results in a form that does not identify the individual.
South Korea	The most relevant legal bases under the PIPA are **consent** and **legitimate interests**. **Consent for secondary purposes**: The organization may rely on the original consent to the extent that the secondary purpose is within the scope of the primary purpose. If the secondary purpose is outside of the scope of the primary purpose, the organization would have to obtain fresh consent pursuant to Article 18 of the PIPA. **Legitimate interests**: Article 15(1)(6) of the PIPA permits organizations to collect and use personal data without the data subject's consent if the collection and use is necessary to achieve a legitimate interest of the organization, and that legitimate interest clearly takes precedence over the data subject's rights. Additional safeguards apply to this provision. The collection and use must be significantly related to the legitimate interest of the organization and must be within a reasonable scope.

Data Protection Principles

DATA MINIMIZATION

Data minimization is a commonly found principle in data protection laws internationally that pertains to the limitation of collection and use of personal data to only what is necessary for a specified purpose.

However, as noted previously, training generative AI models often requires large datasets to learn patterns and generate realistic outputs. These datasets, especially those obtained through "web crawls", may contain significant amounts of personal data which is not strictly necessary for the training of the system, but may be difficult to remove from the dataset.

The principle of data minimization is found in some form in the data protection laws of all 5 jurisdictions covered by this report. It is stated explicitly in the data protection laws of China and South Korea and is implicit in the laws of Australia, Japan, and Singapore.

Jurisdiction	Summary of Relevant Provisions
Australia	The Privacy Act does not expressly recognize the principle of data minimization. However, collection of personal data is subject to a standard of reasonable necessity or relevance. Under APPs 3.2 and 3.3, an organization may only collect personal data if the data is reasonably necessary for, or directly related to, one or more of the organization's functions or activities.
China	Article 6 of the PIPL requires that the collection of personal data must be limited to the smallest scope necessary to achieve the purpose for processing the data. This provision also expressly prohibits excessive collection of personal data.
Japan	The APPI does not expressly recognize the principle of data minimization. However, Article 18(2) of the APPI prohibits businesses from processing personal data beyond the scope necessary to achieve the **purpose of use**, unless they obtain the data subject's consent in advance or satisfy other conditions (see above).
Singapore	The PDPA does not expressly recognize the principle of data minimization. However, Section 18 of the PDPA, which limits collection of personal data for purposes that a reasonable person would find inappropriate, may prevent excessive collection of personal data to some extent.
South Korea	Articles 3(1) and 16(1) of the PIPA require controllers to collect the minimum personal data necessary to fulfill the purpose for processing the data. Further, Article 3(6) of the PIPA requires controllers to minimize the possibility of infringing data subjects' privacy when processing their personal data.

PURPOSE LIMITATION

The principle of purpose limitation requires that personal data can only be collected for specified, explicit, and legitimate purposes and not further processed in a manner incompatible with those purposes.

When training generative AI models on large datasets containing personal data, that personal data may have been collected for other original purposes unrelated to AI training. This discrepancy may complicate compliance with data protection law, as repurposing data for AI training may be deemed incompatible with the initial purpose, potentially requiring additional legal bases, such as consent (see above).

Data protection laws in all 5 of the jurisdictions expressly recognize this principle.

Jurisdiction	Summary of Relevant Provisions
Australia	The Privacy Act implements the principles of purpose limitation in two major respects. Firstly, collection of personal data is limited to purposes which relate to an organization's functions or activities. Under APPs 3.2 and 3.3, an organization may only collect personal data if the data is reasonably necessary for, or directly related to, one or more of the organization's functions or activities. Secondly, an organization may only use or disclose the data for a primary purpose and must satisfy certain conditions before it may use or disclose the data for any other purpose. Specifically, APPs 6.1 and 6.2 require that if an organization holds personal data that was collected for a primary purpose, the organization may only use that data for a secondary purpose if: » the data subject **consents** to the use of their personal data for a secondary purpose, or » the secondary purpose is (directly*) **related to the primary purpose**, if the data subject would **reasonably expect** the organization to use the personal data for the secondary purpose. * For sensitive personal information.
China	Articles 5 and 6 of the PIPL require that processing of personal data must have a clear and reasonable purpose, be directly related to that purpose, and should use a method that has the minimum impact on data subjects' rights and interests.
Japan	The APPI implements the principle of purpose limitation by requiring businesses to identify a **purpose of use** for personal data (Articles 17 and 19). Businesses must obtain the data subject's consent or satisfy other conditions (see above) before using the data for any purpose that is beyond the scope necessary to achieve the purpose of use (Article 18(2)).
Singapore	The PDPA expressly recognizes the principle of purpose limitation. Section 18 of the PDPA only permits organizations to collect, use or disclose personal data about an individual only for purposes that: » a reasonable person would consider appropriate in the circumstances; and » the individual has been informed of, if applicable.
South Korea	Article 3(1) of the PIPA requires data controllers to identify the purpose for processing personal data. Articles 3(2) and 18(1) require controllers to process personal data in an appropriate manner to the extent necessary to fulfill that purpose and not use the data beyond such purposes. Further, Article 3(6) of the PIPA requires controllers to minimize the possibility of infringing data subjects' privacy when processing their personal data.

FAIRNESS

The data protection principle of fairness requires that personal data be processed in a way that is fair and lawful, and respects individual rights.

As discussed in Section 1, when training generative AI models on large datasets, there are risks that the datasets contain biases, inaccuracies, or underrepresentation of certain demographics that may lead these systems to produce output that is biased, discriminatory, or toxic. Such output would certainly contravene the principle of fairness in data protection.

However, in practice, ensuring fairness becomes very complex when using massive datasets, especially those obtained through internet scraping to train generative AI systems. Further, the scale of such datasets may make it challenging to ensure that all data has been collected fairly and lawfully.

Data protection laws in all 5 of the jurisdictions contain some form of requirement that processing of personal data must be fair. While South Korea's PIPA expressly requires fairness in personal data processing, Australia, China, Japan, and Singapore all have implied fairness requirements based on provisions on lawful data collection, good faith, respect for autonomy, and reasonableness standards.

Jurisdiction	Summary of Relevant Provisions
Australia	The Privacy Act does not expressly recognize the principle of fairness. However, under APP 3.5, an organization may only collect personal data by means that are fair and lawful.
China	While the PIPL does expressly recognize the principle of fairness, it only applies this principle to data controllers who provide important internet platform services involving a huge number of users and complicated business types (Article 58). However, the principle of fairness is implicit in other general provisions of the PIPL. Specifically, Article 5 of the PIPL requires that the processing of personal data should be undertaken in good faith and should not involve vitiating factors, such as misrepresentation, fraud, or coercion. Further, Article 6 of the PIPL requires that personal data should be processed in a manner that has the minimum impact on data subjects' rights and interests.
Japan	The APPI does not expressly recognize the principle of fairness. However, Article 3 of the APPI provides a basic principle that personal data should be processed prudently and with respect for the autonomy of data subjects. Further, Article 19 of the APPI prohibits businesses from using personal data in any way that could provoke or induce an unjust act, and Article 20(1) of the APPI prohibits businesses from acquiring personal data by deception or other wrongful means.
Singapore	The PDPA does not expressly recognize the principle of fairness. However, Section 18 of the PDPA subjects purposes for processing personal data to a reasonableness standard. This may serve to prohibit unfair uses of personal data.
South Korea	Article 3(1) of the PIPA requires controllers to collect personal data fairly. Further, Article 3(6) of the PIPA requires controllers to minimize the possibility of infringing data subjects' privacy when processing their personal data.

Personal Data Breaches

Where generative AI models have been trained on datasets containing personal data, these models may generate content that discloses personal data in ways that may cause material or mental harm to data subjects.

For instance, in September 2022, a California-based AI artist found that photographs from her private medical records had been included in a training dataset that was scraped from the internet and had been used to train several image generation models, including Stable Diffusion.[72]

This risk arises from certain features of the transformer architecture which powers many generative AI models today.[73] These models learn by exposure to large datasets and capture the statistical patterns present in the data. In doing so, the model may "**memorize**" information that it was trained on, meaning that the model reproduces specific phrases, sentences, or even longer passages from its training data.[74]

Data protection laws generally require data controllers to secure personal data that is within their control. However, the nature of foundational models raises unique issues, as they may repeat personal data from their training datasets due to the "memorization" issue (see above), either through unintended operation of the system or in response to a malicious prompt that exploits a vulnerability in the AI system. This may lead to unintended disclosure of personal data.

Jurisdiction	Summary of Relevant Provisions
Australia	APP 11 requires organizations to take reasonable steps to: » protect the personal data that they hold from misuse, interference, loss, or unauthorized access, modification, or disclosure; and » proactively delete or de-identify personal data they hold, if data is no longer necessary for any purpose for which it was processed (subject to exceptions for certain legal obligations).
China	The PIPL outlines several operational measures that data controllers must implement to prevent unauthorized access to, breach, tampering or loss of any personal data (Article 51).
Japan	Article 23 of the APPI requires businesses to take necessary and appropriate measures to manage the security of personal data, including preventing leaks, loss, or damage.
Singapore	Section 24 of the PDPA requires organizations to protect personal data in their possession or under their control by, among other provisions, making reasonable arrangements to prevent unauthorized access, collection, use, disclosure, copying, modification or disposal, or similar risks.
South Korea	Article 29 of the PIPA requires controllers to adopt such technical, managerial, and physical measures as are necessary to ensure the safety of personal data and prevent the loss, theft, unauthorized disclosure, forgery, alteration of, or damage to, the data.

Data breaches may also trigger obligations to notify applicable data protection authorities and data subjects. In the latter case, the issue of lack of individualized relationship between the AI operator and the user may create compliance and enforcement difficulties.

Jurisdiction	Summary of Relevant Provisions
Australia	An organization is required to prepare a statement to the OAIC as soon as practicable after discovering an "eligible data breach," (Section 26WK), i.e., an unauthorized access to, or disclosure or loss of personal data, and a reasonable person would conclude that this is likely to result in serious harm to the data subject (Section 26WE). The organization must then notify affected data subjects as soon as practicable after making the statement to the OAIC (Section 26WL). These requirements are subject to exceptions.
China	Article 57 of the PIPL requires organizations to immediately adopt remedial measures and notify the CAC and affected data subjects in the event that a leak, distortion, or loss of personal data has, or might have, occurred. Organizations are permitted not to notify affected data subjects if measures to address the data breach are effective in mitigating harm to data subjects.
Japan	Article 26 of the APPI requires businesses to notify the PPC of any incident involving the security of personal data if the incident is likely to cause harm to the data subject's rights and interests. According to the PPC Order, this notification must be given within 3 to 5 days.[75] Businesses must also "promptly" inform affected data subjects of the breach, unless it is difficult to do so, and the business has implemented necessary measures to protect the data subjects' rights and interests. The PPC has not provided further clarification on the timelines for notifying data subjects.
Singapore	Section 26D of the PDPA requires organizations to notify the PDPC within 3 calendar days of assessing that a data breach has occurred and: » results in, or is likely to result in, significant harm to an affected individual; or » is, or is likely to be, of a significant scale. According to guidelines issued by the PDPC, organizations are expected to complete the above assessment within 30 calendar days (Advisory Guidelines on Key Concepts in the PDPA, paragraph 20.4).[76] Organizations must also notify affected data subjects of the breach in any manner that is reasonable in the circumstances. This requirement is subject to exceptions. In particular, organizations are not required to notify affected data subjects if the organization implemented measures prior to the breach that would render it unlikely that the breach would result in significant harm to the data subject.

Jurisdiction	Summary of Relevant Provisions
South Korea	Article 34 of the PIPA requires controllers to notify the PIPC and affected data subjects "without delay" in the event of a data breach. This requirement does not appear to be subject to exceptions. According to guidelines issued by the PIPC, controllers should notify the PIPC and/or the Korea Internet and Security Agency within 72 hours if the breach involves the personal data of 1,000 or more data subjects, sensitive personal data or unique identifiers, or illegal and unauthorized access to personal data.[77]

Quality of Data

Data protection laws in all 5 of the jurisdictions covered in this Report require organizations to maintain the quality of personal data.

An important consideration in complying with data protection laws in the context of generative AI is that data scraped from the internet may contain personal data that is inaccurate.

Relying on this data and using it for further data processing may conflict with obligations under applicable data protection laws to ensure that personal data is accurate and up to date.[78] For instance, one of the grounds on which the *Garante* temporarily banned ChatGPT in Italy was that it processed inaccurate personal data in violation of Article 5 of the GDPR.[79] One of the *Garante*'s conditions for lifting the temporary ban was that OpenAI provide a tool for data subjects to request rectification or deletion of their data.[80]

Jurisdiction	Summary of Relevant Provisions
Australia	APP 10 requires organizations to take reasonable steps to ensure that: » the personal data that they collect is **accurate**, **up-to-date**, and **complete**; » the personal data that they **use** is accurate, up-to-date, complete, and relevant, having regard to the purpose for which the data will be used.
China	Article 8 of the PIPL requires data controllers to ensure the quality of personal data and avoid adverse impacts on the rights and interests of individuals caused by inaccurate and incomplete personal data.
Japan	Article 22 of the APPI requires businesses to endeavor to keep the content of personal data accurate and up to date, within the scope necessary to achieve the purpose of use.
Singapore	Section 23 of the PDPA requires organizations to make reasonable efforts to ensure that the personal data that they collect is accurate and complete if the organization is likely to use the data to make decisions that affect the data subject or disclose the personal data to another organization.
South Korea	Article 3(3) of the PIPA requires a controller to ensure that personal data is accurate, complete, and up to date to the extent necessary in relation to the purposes for which the personal information is processed.

Rights to Modification and Erasure of Personal Data

All 5 jurisdictions minimally recognize the rights to access and correction of personal data. A further 3 (China, Japan, and South Korea) also provide a right to erasure.

However, giving effect to these rights may be challenging in the context of generative AI.

From a technical perspective, once personal data has been input into generative AI models, effectively managing and tracking its usage becomes a complex, if not challenging, task due to how generative AI systems process information and store/replicate data across various systems.[81]

From a legal compliance perspective, where a generative AI model was trained on a web crawl dataset, it may also be challenging to give effect to the rights of potentially millions (if not billions) of data subjects whose personal data is included in these datasets.

SECTION 3

Summary of Findings and Key Takeaways for APAC

The regulatory landscape for generative AI in APAC is changing at a fast pace. However, based on an analysis of the existing state of generative AI governance in 5 key APAC jurisdictions, this Report has identified some important considerations for policymakers and for deployers and developers of generative AI in the APAC region. Below, we distill the key takeaways for these stakeholders, taking into account the generative AI-specific frameworks, documents, and guidance discussed in Section 1 and detailed in the Appendix.

Takeaways for Policymakers

Takeaway 1: Alignment and interoperability are needed to counter potential policy fragmentation across the region.

A core finding of the Report is that notwithstanding commonalities in certain aspects, **there is a lack of a coordinated approach to generative AI policy** both **within and between** the 5 jurisdictions covered by this Report. This is perhaps unsurprising given the diversity in these 5 jurisdictions (as in the wider APAC region), as well as the lack of mechanisms for supranational coordination compared with other regions, such as Europe.

However, if policymakers continue to develop frameworks to govern generative AI within legislative and national silos, there is a **risk of fragmentation** in the development of these frameworks. Such fragmentation may increase the costs and complexity of compliance across jurisdictions in APAC. This may in turn hinder investment in, and adoption of, potentially valuable technologies at scale, preventing society from reaping the benefits in productivity and innovation from these technologies. It may also create a situation where levels of personal data protection are inconsistent across jurisdictions in the APAC region.

This Report has also identified:

- **A lack of regulatory certainty** (in some areas) **around how existing frameworks apply to AI systems**. The extent to which these laws and rules apply to AI systems is often a matter of legal interpretation, in need of specific regulatory guidance particularly where there are tensions between the nature of processing personal data through Generative AI systems and existing rules.
- **Lack of coordination between legal frameworks (within *and* between jurisdictions)**. Where multiple laws and rules apply to the same issue, there is a risk that their requirements may overlap or even contradict one another. This may create further legal uncertainty, as it may not be clear which laws apply or take precedence in the event of a conflict, or unnecessary layers of regulation.
- **Inconsistency in regulatory responses**. Regulators may not have the same powers to address AI systems that fall within their mandate. This may result in different, and possibly conflicting, responses in different sectors.

It is therefore important for **policymakers to ensure alignment and interoperability** with other leading international frameworks when crafting regulatory responses to generative AI.

Most jurisdictions covered in this Report share the same fundamental aims and have been adopting an incremental approach to AI governance premised on voluntary guidance and consultations. There appears to be an emerging consensus around the risks posed by existing generative AI systems and measures to address them. This emerging consensus could form the basis for regional and international discussions. There are also emerging frameworks at the international level, such as the G7's Hiroshima AI Process Comprehensive Policy Framework, that could aid these discussions.

During FPF's roundtables for this project, several stakeholders emphasized the need for a **common taxonomy** of key terms like "generative AI," "foundation models," and "large language models," that aligns with established regional and global definitions. In this regard, policymakers can benefit from **aligning terminology with emerging international standards** that are being developed in fora including the International Standards Organization (ISO), the Institute of Electrical and Electronics Engineers Standards Association (IEEE SA), the Organization for Economic Cooperation and Development (OECD), and the G7, and encouraging use of standardized terms by all stakeholders. Doing so will aid the establishment of robust standards, guidelines, and regulations for different applications of generative AI.[82]

Takeaway 2: Guidance on the application of existing laws to generative AI should be provided to support legal certainty.

In the absence of AI-specific regulation, **existing technology-neutral laws will continue to be the main source of legal obligations that govern generative AI**. In particular, data protection law plays a major role as all jurisdictions have enacted such laws, and the presence of personal data in training datasets used to train current generative AI models, combined with the consumer-facing nature of many generative AI models, provides data protection authorities with a regulatory lever to govern generative AI.

However, as these laws were not drafted with generative AI in mind and pre-date the current generative AI boom, it would be beneficial if relevant authorities could provide **guidance on how these laws apply to generative AI**. This recommendation is especially relevant to data protection law, as **Section 2 of this Report** has identified several areas of potential ambiguity in how existing data protection laws apply to generative AI systems.

Further, while some DPAs in the 5 jurisdictions have begun issuing guidance on the application of data protection law to certain kinds of AI systems (such as recommendation and decision-making systems[83]), to date, only DPAs in Japan and South Korea have issued guidance on the application of their respective data protection laws to generative AI specifically, and this guidance is still preliminary.

In developing this guidance, and despite the leadership of several data protection authorities, collaboration among relevant regulators within a jurisdiction, including any government body with regulatory authority that may be relevant to generative AI, is essential to ensure that each jurisdiction's approach to generative AI governance is consistent across regulatory domains and avoids the risk of regulatory fragmentation within that jurisdiction. Ideally, relevant regulators should coordinate on priorities and approaches and work together to identify potential gaps in existing frameworks, proposing targeted reforms as necessary.

In particular, it may be helpful to look not only for gaps in existing frameworks but also overlaps where multiple legislative or regulatory frameworks may govern the same issue. Such overlaps may complicate compliance, especially if requirements are contradictory.

Takeaways for Industry including Developers and Deployers of Generative AI Systems

As of early 2024, there is an emerging body of voluntary guidance issued from the 5 APAC jurisdictions studied, outlining good practices that developers and deployers of generative AI could consider adopting in their approaches to govern this technology.

This subsection of the Report summarizes these practices, building on the commonalities identified in **Section 1** and serving as a resource for developers and deployers of generative AI systems thinking through generative AI governance in APAC or comparing existing approaches in APAC with legally binding requirements in the EU and US.

Takeaway 3: All five jurisdictions recognize developing internal AI governance and risk management policies as a good practice.

As shown from our survey of early regulatory responses to generative AI outlined in Section 1, policymakers in APAC have highlighted that a good practice before developing or deploying a generative AI system is to design a robust internal AI policy and strategy to encourage the organization to foster a culture of responsible innovation.

Existing AI governance frameworks in the 5 jurisdictions point to the following as relevant factors to consider:

- Assessing the organization's AI proficiency.[84]
- Setting governance principles and goals.[85]
- Integrating ethical guidelines, risk management protocols, and compliance measures.[86]
- Ensuring compliance with existing laws and guidelines.[87]
- Conducting risk and impact assessments to systematically evaluate potential harms to guide mitigation efforts.[88]
- Documenting risk and impact assessments to facilitate transparency and build organizational accountability.[89]
- Clearly allocating responsibilities within the organization, including potentially establishing an AI taskforce or committee to coordinate efforts.[90]
- Training employees in the design, function and implementation of AI systems.[91]
- Regularly reviewing governance structures and measures to ensure alignment with objectives and address evolving risks.[92]

Takeaway 4: Effective governance is essential to mitigate model bias and discriminatory outputs from generative AI systems.

When developing generative AI systems, organizations must prioritize data governance, including implementing good data practices, evaluating training data sources, and evaluating model output for representativeness. Identifying potential biases in the model is key to mitigate against the risk of discriminatory or harmful output.

Existing AI governance frameworks in the 5 jurisdictions have highlighted the following potential measures to mitigate bias and prevent or discriminatory outputs from generative AI systems:

- Thoroughly evaluating training data sources for representativeness and potential biases.[93]
- Documenting data provenance to enable traceability and accountability.[94]
- Regularly auditing data quality across dimensions like accuracy, completeness, and relevance.[95]
- Proactively conducting bias assessments and ethical reviews of training data.[96]
- Moderating and redacting problematic content from training data.[97]
- Fine-tuning models after initial training to reduce harmful outputs.[98]
- Employ output filtering techniques to catch and block biased generations.[99]
- Leverage bias detection tools during data preprocessing.[100]
- Continuously monitoring and updating datasets with human oversight.[101]

Takeaway 5: Ensuring privacy by design in the development and deployment of generative AI systems can build public trust.

When developing underlying generative AI models and deploying generative AI-based systems and applications, organizations can benefit from adopting a "Privacy by Design" approach that builds in data protection safeguards from the earliest stages and at regular intervals thereafter. This can help build public trust in the technology.

Potential privacy-preserving measures identified or recommended in generative AI-specific policy documents across the five jurisdictions include:

- Minimizing collection and use of personal data.[102]
- Redacting or anonymizing personal data in training datasets.[103]
- Conducting Data Protection Impact Assessments (DPIAs).[104]
- Ensuring legal compliance for data collection and usage.[105]
- Developing and publishing a privacy policy to address the organization's use of personal data in training an AI model.[106]
- Obtain informed consent from users and provide mechanisms for users to opt out of collection or use of their personal data to train generative AI models.[107]

A limited yet important privacy enhancing technique that can offer an alternative to personal data to train generative AI models is the use of **synthetic data**: artificial data generated from original data by an AI model that has been trained to reproduce the characteristics and structure of the original data.[108]

Such data can potentially be used for pre-training, fine-tuning, and testing AI models,[109] and preliminary research has found that models trained on synthetic data achieved over 90% of the quality of models trained on real datasets.[110]

According to the Confederation of European Data Protection Organisations (CEDPO), potential benefits of synthetic data include enhanced privacy by minimizing the use of personal data, better data quality through "near-perfect" labeling, reduced costs, and reduced cybersecurity attack surfaces. However, synthetic data is not synonymous with anonymous data and carries a risk of reidentification.[111] The use of other Privacy Enhancing Technologies, such as differential privacy, in combination with synthetic data, could mitigate the risks of reidentification but may not completely remove it.[112]

Takeaway 6: Implementing safety and security measures is paramount for safer generative AI systems.

Implementing appropriate safety and security measures helps to ensure that generative AI systems are used safely and responsibly, minimizing the potential for misuse or harm to users and third parties.

Existing AI governance frameworks in the 5 jurisdictions have highlighted the following potential safety measures:

- Conducting risk and impact assessments, prompt testing and design, and ongoing evaluation.[113]
- Adding friction points, such as educative prompts or inappropriate content detection, when users attempt to generate content.[114]
- Implementing age-appropriate design with effective age verification measures, limiting content generation for underage users to age-appropriate material.[115]
- Implementing policies and processes to detect malicious actors or harmful content, testing models for potential misuse and putting safeguards in place to prevent harmful content generation.[116]

They also highlight the following potential security measures:

- Engaging in thorough testing and evaluation processes to mitigate risks.[117] This could include voluntary certifications, audits, and third-party assessments,[118] as well as "crowdsourcing" the detection of vulnerabilities in open-source models.[119]
  - Testing and evaluation processes could also include "**red teaming**"[120] – a practice where an authorized security team pretends to be attackers and tries to break into an organization's systems to test its security.[121]
- Establishing channels to share information regarding risks and best practices, including incident reporting.[122]
  - This includes, but is not limited to, complying with notifiable data breach requirements in data protection laws (see above).

Takeaway 7: All five jurisdictions recognize that providing meaningful transparency in the development and deployment of generative AI systems is essential.

As discussed in Section 1, transparency is a multi-faceted concept that is closely related to accountability. In particular, it allows scrutiny of potential harms, calibrates user expectations, and ultimately nurtures public trust in generative AI technologies.

Existing AI governance frameworks in the 5 jurisdictions have highlighted the following potential measures to facilitate meaningful transparency in the development and deployment of generative AI systems:

- Publishing clear organizational policies covering user safety, privacy, terms of use, content guidelines, and impact assessments.[123]
- Providing notices on data collection purposes, factual inaccuracies, and dissuading sharing of personal and/or confidential information.[124]
- Enhancing context-appropriate explainability and interpretability to clarify how models function and arrive at outputs. This could include:
  - Maintaining comprehensive documentation on data provenance, design choices, training procedures, performance metrics, and ethical evaluations.[125]
  - Utilizing model cards, system cards, and value alignment cards to present technical details in an accessible manner.[126]
  - Clarifying models' capabilities, limitations, and intended or prohibited uses.[127]
- Disclosing transparency reports.[128]
- Providing mechanisms for stakeholders to request further information, provide feedback, and seek redress.[129]

While full transparency may be impossible, an important consideration is ensuring that appropriate explanations are tailored to the needs of different stakeholders, which may include regulators, downstream providers of services that use generative AI models, and end-users.[130]

In addition, technical explanations of algorithms may not be useful to the general public. It may be helpful instead to focus on real-world impacts rather than solely technical inner workings to improve meaningful consent.

Takeaway 8: Indicating that content is AI-generated and enabling traceability are unanimously included in the generative AI frameworks studied.

Policymakers in the 5 jurisdictions were unanimous in highlighting the need for mechanisms to enable stakeholders, including regulators and the general public, to identify content as AI-generated.[131] This is closely related to transparency but also has implications for safety and security.

Industry is already working on technology to embed digital labels or watermarks in AI-generated content indicating that the content was generated by their system. For instance, the Coalition for Content Provenance and Authenticity (C2PA) is developing an open industry standard using cryptography to embed digital signatures and ownership details into AI-generated content.[132]

However, solutions like these may not be suitable for AI-generated text, since text can be more easily separated from metadata. While statistical watermarking and other techniques for text are emerging,[133] this area is still in early development.[134]

Noting that the technology to accomplish this is still at an early stage of development, organizations could consider implementing measures to make AI-generated outputs detectable, such as:

- Digital labeling or watermarking indicating AI-generated provenance, whether visible markers or embedded metadata.[135]
- Exploring statistical watermarking techniques tailored for text data.[136]
- Coordinating efforts to imprint subtle "fingerprints" in training data or model architectures that enable detection of AI-generated output.

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**[137] 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.[138] 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.[139] To guide the public consultation, the DISR released a discussion paper, titled "**Safe and Responsible AI in Australia**."[140]

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.[141]

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 released[142] its "**Tech Trends Position Statement on Generative AI**."[143] 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,[144] 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: » **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: » **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)[145] – which includes the OAIC as well as Australian Competition and Consumer Commission, the Australian Communications and Media Authority, and the eSafety Commissioner – released[146] a working paper examining LLMs and their impact on the regulatory roles of each member of the DP-REG.[147]

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).[148]

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.[149]
- "**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.[150]

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

Principle	Elaboration
Advancement of human welfare (增进人类福祉)	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. 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.
Promotion of fairness and justice (促进公平公正)	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. When providing AI products and services, AI actors should fully respect and help vulnerable groups and special groups, and provide appropriate alternatives as necessary.

Principle	Elaboration
Protection of privacy and security (保护隐私安全)	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. AI actors should process personal data according to the principles of legality, propriety, necessity, and good faith, and guarantee personal privacy and data security. 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.
Assurance of controllability and trustworthiness (确保可控可信)	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. AI actors should also ensure that AI is always under human control.
Strengthening accountability (强化责任担当)	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. AI actors should also establish AI accountability mechanisms and should not avoid investigations into responsibility or evade their own responsibilities.
Improving ethical literacy (提升伦理素养)	AI actors should actively learn about and spread awareness of AI ethics. AI actors should objectively understand ethical issues and should not underestimate or exaggerate ethical risks. AI actors should actively carry out or participate in discussions of AI ethical issues. AI actors should thoroughly promote the practice of AI ethical governance and improve their ability to respond to ethical issues.

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**	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. 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.
	Actively practicing ethics and demonstrating how to put ethics into practice	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. 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.
	Correctly exercising authority	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. 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. They should also prohibit improper exercises of authority that may harm the legitimate rights and interests of natural persons, legal persons, and other organizations.
	Strengthening risk prevention	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. They should also establish effective warning mechanisms and improve their capabilities to control and handle ethical risks.
	Promoting inclusive openness	Persons involved in the management of AI-related activities should give full consideration to the rights and expectations of all stakeholders in AI. They should encourage the application of diversified AI technologies to address practical economic and social development issues. 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.

Activity	Responsibility	Elaboration
Research and development	**Strengthening "self-discipline"**	Persons involved in researching and developing AI should exercise self-restraint in AI-related research and development. 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.
	Improving data quality	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. They should also improve the integrity, timeliness, consistency, standardization, and accuracy of data.
	Enhancing security and transparency	Across the algorithm design, implementation, and application stages, persons involved in researching and developing AI should: » 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	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. They should also avoid potential data and algorithm biases, and strive to achieve inclusiveness, fairness, and non-discrimination in AI systems.

Activity	Responsibility	Elaboration
Supply	**Respecting market rules**	Persons involved in supplying AI-related products and services should strictly comply with regulations governing market access, competition, and transactions. 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.
	Strengthening quality control	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. They should not operate, sell, or provide products and services that do not meet quality standards.
	Safeguarding users' rights and interests	Persons involved in supplying AI-related products and services should clearly inform users, identifying the functions and limitations of the products and services. 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. They also should not create barriers to the equal use of AI by users.
	Strengthening emergency support	Persons involved in supplying AI-related products and services should research and formulate emergency mechanisms and plans or measures to compensate users for losses. 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.

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)[151] (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: » 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: » 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)[152] 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**, » 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: » 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: » 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)[153] – 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."[154] 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	Data from user prompts should only be used to train a model if users have authorized such use. 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.
Annotating training data	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.
Using models developed by third parties	Service providers who use models developed by third parties should use models that have been filed with the competent authority.
Safety of generated content	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. In each conversation, the input information from users should undergo safety checks to guide the model to generate positive and constructive content. 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. Technical measures should also be adopted to improve the accuracy and reliability of generated content.
Use of generative AI in certain sectors	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.
Minors	For services applicable to minors: » 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. Technical or managerial measures should be taken to prevent minors from using services not applicable to minors.
Transparency	For services provided through interactive interfaces, information should be provided about: » the applicable users, scenarios, purposes; » The limitations of the service; and » A summary of the generative AI model or algorithm used. Information on whether user inputs are used to train the model, and how to opt out of this, should be prominently displayed.
Supply chains	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. 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.

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	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; 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. 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.
Model updates and upgrades	Security management strategies should be formulated for model updates and upgrades. A management mechanism should be established to organize security assessments again after significant model updates or upgrades.
Service stability and continuity	The training environment should be isolated from the inference environment to prevent data leakage and unauthorized access. Continuous monitoring of model input content should be conducted to prevent malicious input attacks, such as DDoS, XSS, injection attacks, etc. 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. Backup mechanisms and recovery strategies for data, models, frameworks, tools, etc., should be established, with a focus on ensuring business continuity.

Draft AI Law (March 2024)

On 31 May 2023, China's State Council released its legislative work plan for 2023.[155] 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.[156] 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)[157] 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:

Social Principles of AI	Elaboration
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)[158] 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.

Stage	Action Targets
Conditions and Risk Analysis	**1-1 Understanding positive and negative impacts of using AI**. Developers and operators should understand not only positive impacts but also negative impacts that AI systems may have, including unintended risks. 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.
	1-2 Understanding social acceptance of the use of AI. 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. In addition, even after the full-scale operation, companies should obtain opinions of stakeholders again and update their perspectives in a timely manner.
	1-3 Understanding the company's AI proficiency. Developers and operators should evaluate and re-evaluate in a timely manner their AI proficiency based on: » 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, except in situations where a company assesses that negative impacts of their AI system are minor. 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.

Stage	Action Targets
Goal Setting	**2-1 Considering and setting AI governance goals**. Developers and operators should consider whether or not to set their own AI governance goals based on the Social Principles of Human-Centric AI. 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.
Designing an AI Management System to Achieve AI Governance Goals This includes both technological and organizational systems.	**3-1 Employing "gap analysis" between AI governance goals and the current state of AI governance and addressing gaps**. 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. Companies should provide sufficient information about the gaps and measures to address the gaps, as well as make a contact point easily accessible. 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. Developers should acquire data sets from data providers that provide sufficient information.
	3-2 Improving the literacy of AI management personnel. 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. 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.
	3-3 Reinforcing AI management through cooperation between companies. 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. 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. 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.
	3-4 Preventing and responding to incidents. 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. 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. Further, they should consider conducting rehearsal exercises relevant to such guidelines and plans, as appropriate.

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.[159]

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.[160] METI released the final version of the Guidelines in April 2024.[161]

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: » 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: » 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: » 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.[162]

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	**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. Examples of roles and responsibilities that could be allocated include: » 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.
	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. Examples of possible measures that could be implemented include: » 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	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**. 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.
Operations management	The Framework outlines **good data accountability practices** for training datasets used to train AI models. These include: » 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.
	The Framework also outlines possible measures to ensure that the AI model **makes decisions that are explainable, repeatable, robust, reproducible, and auditable**. These include assessment and testing, as well as regular model tuning to respond to changes over time.

Section	Subsections
Stakeholder interaction and communication	The Framework discusses different strategies for organizations to **build trust with stakeholders**. These include: » 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. It also discusses relevant factors, such as audience, purpose, and context.
	The Framework discusses potential measures that organizations could consider implementing to **manage relationships with their customers**, such as: » 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**."[163]

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.[164] Note that AI Verify currently does not apply to generative AI systems, including LLMs.[165]

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 A 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.[166]

To guide the sandbox, the IMDA also released a draft catalog of current benchmarks and methods to evaluate LLMs, titled "**Cataloguing LLM Evaluations**,"[167] 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**"[168] 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	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. 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. 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. 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**.
Data	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. **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.
	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.
	The Proposed Framework also suggests that stakeholders should consider **expanding the available pool of trusted reference datasets** for model development, benchmarking, and evaluation. It highlights that **governments** could play a role in curating repositories of representative training data sets for their specific cultural, social, or linguistic contexts.
Trusted Development and Deployment	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.
	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. 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.
	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. It recommends that industry and sectoral policy makers jointly improve evaluation benchmarks and tools, while still maintaining coherence between baseline and sector specific requirements.

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: » **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)[169] were released on December 23, 2020, at the 19th meeting of the Presidential Committee on the 4th Industrial Revolution.[170]

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.[171]

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)	Solidarity should be maintained in relationships between various groups, and AI should be used with sufficient consideration for future generations. Fair opportunities should be ensured for diverse stakeholders throughout the entire lifecycle of AI. Efforts should be made for international cooperation in the development and utilization of ethical AI.
Data Management (데이터 관리)	Individual data, including personal data, should be used in line with its intended purpose, and not for any other purpose. Data quality and risk must be managed to minimize data bias throughout the entire process of data collection and use.
Accountability (책임성)	Efforts should be made to minimize damage that may occur by establishing a responsible entity in the process of developing and using AI. Responsibilities between AI design and developers, service providers, and users must be clearly specified.
Safety (안전성)	Efforts must be made to prevent potential risks and ensure safety throughout the entire process of developing and use of AI. 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.
Transparency (투명성)	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. 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.

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**,"[172] 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.[173]

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.[174] 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).[175] 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**: » 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: » 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**,"[176] 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."[177]

According to the ministers' statement, the Framework 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;[178]
- the "**International Guiding Principles for Organizations Developing Advanced AI Systems**" (International Guiding Principles)[179]
- the voluntary "**International Code of Conduct for Organizations Developing Advanced AI Systems**" (International Code of Conduct)[180] 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.	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. The Code of Conduct highlights several risks that should be considered in designing and implementing testing measures: » 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. Organizations commit to work in collaboration with relevant actors across sectors, to assess and adopt mitigation measures to address these risks, in particular systemic risks. Organizations making these commitments should also endeavor to advance research and investment on the security, safety, bias and disinformation, fairness, explainability and interpretability, and transparency of advanced AI systems and on increasing robustness and trustworthiness of advanced AI systems against misuse. These measures should be documented and supported by regularly updated technical documentation.

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 **r**eport 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.	This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems. These reports, instruction for use, and relevant technical documentation, as appropriate, should be kept up-to-date and should include, for example; » Details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights, » Capacities of a model/system and significant limitations in performance that have implications for the domains of appropriate use, » Discussion and assessment of the model's or system's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness, and » The results of red-teaming conducted to evaluate the model's/system's fitness for moving beyond the development stage. Robust documentation processes include technical documentation and instructions for use.

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.	This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle. Organizations should establish or join mechanisms to develop, advance, and adopt, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems. This should also include ensuring appropriate and relevant documentation and transparency across the AI lifecycle in particular for advanced AI systems that cause significant risks to safety and society. Organizations should collaborate with other organizations across the AI lifecycle to share and report relevant information to the public with a view to advancing safety, security, and trustworthiness of advanced AI systems. Organizations should also collaborate and share the aforementioned information with relevant public authorities, as appropriate. Such reporting should safeguard intellectual property rights.
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.	Organizations should put in place appropriate organizational mechanisms to develop, disclose, and implement risk management and governance policies, including for example accountability and governance processes to identify, assess, prevent, and address risks, where feasible throughout the AI lifecycle. This includes disclosing where appropriate privacy policies, including for personal data, user prompts, and advanced AI system outputs. Organizations are expected to establish and disclose their AI governance policies and organizational mechanisms to implement these policies in accordance with a risk-based approach. This should include accountability and governance processes to evaluate and mitigate risks, where feasible throughout the AI lifecycle. The risk management policies should be developed in accordance with a risk-based approach and apply a risk management framework across the AI lifecycle as appropriate and relevant, to address the range of risks associated with AI systems, and policies should also be regularly updated. Organizations should establish policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices.

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.	These may include securing model weights and algorithms, servers, and datasets, such as through operational security measures for information security and appropriate cyber/ physical access controls.	This also includes performing an assessment of cybersecurity risks and implementing cybersecurity policies and adequate technical and institutional solutions to ensure that the cybersecurity of advanced AI systems is appropriate to the relevant circumstances and the risks involved. Organizations should also have in place measures to require storing and working with the model weights of advanced AI systems in an appropriately secure environment with limited access to reduce both the risk of unsanctioned release and the risk of unauthorized access. This includes a commitment to have in place a vulnerability management process and to regularly review security measures to ensure they are maintained to a high standard and remain suitable to address risks. This further includes establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets, for example, by limiting access to proprietary and unreleased model weights.
Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content.	This includes, where appropriate and technically feasible, content authentication 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. 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.	Organizations should collaborate and invest in research, as appropriate, to advance the state of the field.

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").[181]

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,[182] 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.[183]

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.[184]

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.[185] 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.[186] 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.[187]

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.[188]

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 Parliament[189] 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).

ENDNOTES

1 See, for example, *The Spectrum of Artificial Intelligence – An Infographic Tool* (December 14, 2020), available at https://fpf.org/blog/the-spectrum-of-artificial-intelligence-an-infographic-tool/, *The Spectrum of AI – Companion to the FPF AI Infographic* (updated July 18, 2023), available at https://fpf.org/blog/newly-updated-report-the-spectrum-of-artificial-intelligence-companion-to-the-fpf-ai-infographic/, and *Generative AI for Organizational Use: Internal Policy Checklist* (July 13, 2023), available at https://fpf.org/resource/fpf-releases-generative-ai-internal-policy-checklist-to-guide-development-of-policies-to-promote-responsible-employee-use-of-generative-ai-tools/

2 The **Appendix to this Report** includes brief summaries of the EU's AI Act and the US' Executive Order on the Safe, Secure, and Trustworthy Development of Artificial Intelligence.

3 In November 2022, OpenAI launched ChatGPT, a consumer-facing generative AI tool that can create predictive content based on natural language input. Since then, a number of competitors have emerged in the space in both the business-to-consumer and business-to-business markets, as well as a number of more sophisticated generative AI tools.

4 https://www.ibm.com/topics/ai-model

5 Jebara, T. (2012). *Machine learning: discriminative and generative* (Vol. 755). Springer Science & Business Media.

6 Weidinger, L. et al. (2021), Ethical and social risks of harm from Language Models, https://arxiv.org/abs/2112.04359, page 22.

7 https://openai.com/sora

8 Vaswani, A. et al. (2017), *Attention Is All You Need*, https://arxiv.org/abs/1706.03762.

9 Triguero, I, et al. (2024). "General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance." *Information Fusion* 103 (2024): 102135.

10 https://openai.com/blog/openai-codex

11 https://deepmind.google/discover/blog/competitive-programming-with-alphacode/

12 https://research.ibm.com/blog/ai-for-code-project-wisdom-red-hat

13 Chithrananda, S. et al. (2020) *ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction*. https://arxiv.org/abs/2010.09885

14 Irwin, R. et al. (2022) "Chemformer: a pre-trained transformer for computational chemistry." *Machine Learning: Science and Technology* 3(1). https://iopscience.iop.org/article/10.1088/2632-2153/ac3ffb

15 https://research.ibm.com/blog/molecular-transformer-discovery

16 https://www.microsoft.com/en-us/research/group/autonomous-systems-group-robotics/articles/introducing-climax-the-first-foundation-model-for-weather-and-climate/

17 https://newsroom.ibm.com/2023-08-03-IBM-and-NASA-Open-Source-Largest-Geospatial-AI-Foundation-Model-on-Hugging-Face

18 https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/

19 Assael, Y. et al. (2022). "Restoring and attributing ancient texts using deep neural networks." *Nature* 603, 280–283. https://www.nature.com/articles/s41586-022-04448-z

20 Al Quraishi, M. (2021). "Machine learning in protein structure prediction," *Current Opinion in Chemical Biology* 65 1–8, https://doi.org/10.1016/J.CBPA.2021.04.005; Rives, A. et al. (2021). "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.: *Proceedings of the National Academy of Sciences* 118, 15 (2021). https://doi.org/10.1073/pnas.2016239118.

21 Rothchild, A. et al. (2021). *C5T5: Controllable Generation of Organic Molecules with Transformers*. https://arxiv.org/abs/2108.10307

22 https://openai.com/blog/chatgpt

23 Mollick, E. (2022), "ChatGPT is a Tipping Point for AI" *Harvard Business Review.* https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai

24 https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework

25 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/

26 https://www8.cao.go.jp/cstp/ai/ningen/ningen.html. English translation available at https://www8.cao.go.jp/cstp/ai/humancentricai.pdf

27 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

28 https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework

29 Available at https://openresearch-repository.anu.edu.au/bitstream/1885/277585/1/SKAI_31.pdf. No authoritative English language translation is available.

30 https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_Y2B1M0R6G2I2P1B0V2X9H4Z0X3M3J2

31 https://www.chiefscientist.gov.au/sites/default/files/2023-06/Rapid%20Response%20Information%20Report%20-%20Generative%20AI%20v1_1.pdf

32 https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai

33 https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/Safe-and-responsible-AI-in-Australia-discussion-paper.pdf

34 https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf

35 https://dp-reg.gov.au/publications/working-paper-2-examination-technology-large-language-models

36 http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm

37 http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

38 https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf

39 https://www.ppc.go.jp/files/pdf/230602_kouhou_houdou.pdf

40 https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html

41 https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

42 https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf

43 https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9055#LINK

44 https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9083

45 Commonly referred to as "hallucinations."

46 *Supra* n 6.

47 *Ibid*.

48 https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

49 Tirumala, K. et al. (2022). "Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models" *36th Conference on Neural Information Processing Systems* (NeurIPS 2022). https://arxiv.org/pdf/2205.10770.pdf

50 *Supra* n 6, page 19.

51 Burgess, M. (2023). "The Hacking of ChatGPT Is Just Getting Started" *Wired* https://www.wired.com/story/chatgpt-jailbreak-generative-ai-hacking/; Oremus, W. (2023). "The clever trick that turns ChatGPT into its evil twin" *The Washington Post.* https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/

52 *Supra* n 23.

53 OECD (2023), "AI language models: Technological, socio-economic and policy considerations", *OECD Digital Economy Papers*, No. 352, OECD Publishing, Paris, https://doi.org/10.1787/13d38f92-en, page 34.
54 *Ibid*.
55 *Supra* n 6, page 15.
56 https://digi.org.au/disinformation-code
57 https://www.legislation.gov.au/Latest/C2022C00361
58 https://www.gov.cn/xinwen/2021-08/20/content_5632486.htm. A nonbinding English translation is available at http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm.
59 https://elaws.e-gov.go.jp/document?lawid=415AC0000000057. An unofficial English translation is available at https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en
60 https://sso.agc.gov.sg/Act/PDPA2012?
61 https://www.law.go.kr/법령/개인정보보호법 An unofficial English translation is available at https://elaw.klri.re.kr/kor_service/lawView.do?hseq=53044&lang=ENG
62 Zanfir-Fortuna, G. (2023) *How Data Protection Authorities are De Facto Regulating Generative AI,* https://fpf.org/blog/how-data-protection-authorities-are-de-facto-regulating-generative-ai/.
63 Paulger, D. (2022). *Balancing Organizational Accountability and Privacy Self-Management in Asia-Pacific.* https://fpf.org/blog/new-report-promotes-accountability-based-approach-to-data-protection-in-the-apac-region/
64 https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/
65 https://commoncrawl.org/overview
66 https://paperswithcode.com/dataset/c4
67 https://laion.ai/blog/laion-400-open-dataset/
68 https://laion.ai/blog/laion-5b/
69 Verma, P. and Oreumus, W. (2023). "ChatGPT invented a sexual harassment scandal and named a real law prof as the accused." *The Washington Post.* https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/
70 Kaye, B. (2023). "Australian mayor readies world's first defamation lawsuit over ChatGPT content" *Reuters.* https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/
71 https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/ag-on-key-concepts/advisory-guidelines-on-key-concepts-in-the-pdpa-17-may-2022.pdf
72 Edwards, B. (2023) "Artist finds private medical record photos in popular AI training data set." *Ars Technica.* https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/
73 *Supra* n 6, page 20.
74 *Supra* n 49.
75 https://www.ppc.go.jp/personalinfo/legal/leakAction/
76 https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/ag-on-key-concepts/advisory-guidelines-on-key-concepts-in-the-pdpa-17-may-2022.pdf
77 https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS217&mCode=D010030000&nttId=10059
78 Confederation of European Data Protection Organisations, (2023). *Generative AI: The Data Protection Implications*, https://cedpo.eu/generative-ai-the-data-protection-implications/, pages 4-5.
79 https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english
80 https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9874702#english
81 *Supra* n 78, page 6.
82 World Economic Forum (2023). *The Presidio Recommendations on Responsible Generative AI*. https://www.weforum.org/publications/the-presidio-recommendations-on-responsible-generative-ai/, page 3.
83 See, for example, the Personal Data Protection Authority of Singapore's Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems, available at https://www.pdpc.gov.sg/guidelines-and-consultation/2024/02/advisory-guidelines-on-use-of-personal-data-in-ai-recommendation-and-decision-systems
84 **China**: Ethical Principles for New Generation AI, Section II. **Japan**: Governance Guidelines for Implementation of AI Principles (Ver 1.1), page 15.
85 **Australia**: eSafety Commissioner Position Statement, pages 6, 29. **China**: Ethical Principles for New Generation AI, Section II. **Japan**: Governance Guidelines for Implementation of AI Principles (Ver 1.1), page 17. **Singapore**: Model AI Governance Framework, page 17. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
86 **China**: Ethical Principles for New Generation AI, Section III.
87 **China**: Ethical Principles for New Generation AI, Sections II, III; Deep Synthesis Regulations, Articles 4-6; Interim Generative AI Measures, Article 7. **Japan**: Governance Guidelines for Implementation of AI Principles (Ver 1.1), page 17. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
88 **Australia**: Rapid Research Information Report, page 12; Safe and Responsible AI in Australia Discussion Paper, pages 40-41; eSafety Commissioner Position Statement, pages 6, 29. **China**: Ethical Principles for New Generation AI, Section II. **Japan**: Governance Guidelines for Implementation of AI Principles (Ver 1.1), page 9; Guidelines for AI Business Operators, Recommendation D-2. **Singapore**: Model AI Governance Framework, page 29. **South Korea:** Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
89 **Australia**: Safe and Responsible AI in Australia Discussion Paper, pages 40-41. **Singapore**: Model AI Governance Framework, page 29.
90 **Australia**: Government's Interim Response to Safe and Responsible AI in Australia Consultation, page 20; eSafety Commissioner Position Statement, page 29. **China**: Ethical Principles for New Generation AI, Section II; Deep Synthesis Regulations, Articles 7-8. **Singapore**: Model AI Governance Framework, pages 21-23.
91 **Australia**: Safe and Responsible AI in Australia Discussion Paper, pages 6, 40-41; Government's Interim Response to Safe and Responsible AI in Australia Consultation, page 20; eSafety Commissioner Position Statement, page 8. **Japan**: Governance Guidelines for Implementation of AI Principles (Ver 1.1), page 27.
92 **Singapore**: Model AI Governance Framework, page 29.
93 **Australia**: eSafety Commissioner Position Statement, page 6. **Japan**: Guidelines for AI Business Operators, Recommendation D-2, D-3. **Singapore**: Model AI Governance Framework, page 36.
94 **Singapore**: Model AI Governance Framework, pages 37-38.
95 **China**: Basic Security Requirements, Section 5. **Japan**: Guidelines for AI Business Operators, Recommendation D-2, D-3. **Singapore**: Model AI Governance Framework, pages 37-38.
96 **Australia**: eSafety Commissioner Position Statement, page 6. **Japan**: Guidelines for AI Business Operators, Recommendation D-2, D-3. **Singapore**: Model AI Governance Framework, page 24.
97 **China**: Basic Security Requirements, Section 5.
98 **Australia**: eSafety Commissioner Position Statement, page 7. **China**: Basic Security Requirements, Section 6. **Japan**: Guidelines for AI Business Operators, Recommendation P-3.

99 **Australia**: eSafety Commissioner Position Statement, pages 7-8. **Japan**: Guidelines for AI Business Operators, Recommendation P-3.
100 **Australia**: eSafety Commissioner Position Statement, page 6.
101 **Singapore**: Model AI Governance Framework, page 40.
102 **Australia**: eSafety Commissioner Position Statement, page 6. **China**: Interim Generative AI Measures, Article 11. **South Korea:** Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
103 **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
104 **Australia**: Government's Interim Response to Safe and Responsible AI in Australia Consultation, page 20. **Japan**: Guidelines for AI Business Operators, Appendix 3, Section 4. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
105 **China**: Interim Generative AI Measures, Articles 7, 11.
106 **China**: Basic Security Requirements, Section 7. **Japan**: Guidelines for AI Business Operators, Recommendations D-2, P-4. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
107 **Australia**: Rapid Research Information Report, page 12; eSafety Commissioner Position Statement, pages 18, 30-32. **China**: Deep Synthesis Regulations, Article 14; Interim Generative AI Measures, Articles 7, 11; Basic Security Requirements, Section 7. **Singapore**: Model AI Governance Framework, page 56. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
108 *Supra* n 78, page 21-22. See also UK Information Commissioner's Office, *G7 DPAs' Emerging Technologies Working Group use case study on privacy enhancing technologies*, available at https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/case-studies/g7-dpas-emerging-technologies-working-group-use-case-study-on-privacy-enhancing-technologies/
109 Competition and Markets Authority (UK), (2023). *AI Foundation Models: Initial Report*, https://www.gov.uk/government/publications/ai-foundation-models-initial-report, page 32.
110 https://lmsys.org/blog/2023-03-30-vicuna/
111 CEDPO, Generative AI: The Data Protection Implications (16 October 2023), page 21-22.
112 *Supra* n 109, page 33. See also UK ICO, G7 DPAs' Emerging Technologies Working Group use case study on privacy enhancing technologies,
113 **Australia**: eSafety Commissioner Position Statement, page 8. **China**: Deep Synthesis Regulations, Article 10; Interim Generative AI Measures, Article 17; Basic Security Requirements, Section 9. **Japan**: Guidelines for AI Business Operators, Recommendation D-2, D-5. P-2, P-5. **Singapore**: AI Verify Discussion Paper, page 22-24; Proposed Model Governance Framework for Generative AI, page 10.
114 **Australia**: eSafety Commissioner Position Statement, page 8.
115 **Australia**: eSafety Commissioner Position Statement, page 29. **China**: Interim Generative AI Measures, Article 10; Basic Security Requirements, Section 7.
116 **Australia**: eSafety Commissioner Position Statement, page 8. **China**: Deep Synthesis Regulations, Article 10; Interim Generative AI Measures, Article 14; Basic Security Requirements, Section 7. **Singapore**: Proposed Model Governance Framework for Generative AI, page 16.
117 **Australia**: Government's Interim Response to Safe and Responsible AI in Australia Consultation, pages 13, 20; eSafety Commissioner Position Statement, pages 8, 29. **China**: Deep Synthesis Regulations, Articles 10, 15; Basic Security Requirements, Section 6. **Japan**: Guidelines for AI Business Operators, Recommendation D-2, D-5. P-2, P-5. **Singapore**: AI Verify Discussion Paper, page 22-24; Proposed Model Governance Framework for Generative AI, pages 11-12.
118 **Australia:** Safe and Responsible AI in Australia Discussion Paper, pages 40-41. **Singapore**: AI Verify Discussion Paper, page 24; Proposed Model Governance Framework for Generative AI, pages 15-16.
119 **Singapore**: AI Verify Discussion Paper, page 24.
120 **Australia**: eSafety Commissioner Position Statement, pages 8, 29. **Singapore**: AI Verify Discussion Paper, page 25; Proposed Model Governance Framework for Generative AI, page 11.
121 https://csrc.nist.gov/glossary/term/red_team
122 **Australia**: Government's Interim Response to Safe and Responsible AI in Australia Consultation, page 20. **China**: Deep Synthesis Regulations, Article 10. **Japan**: Guidelines for AI Business Operators, Recommendation D-5. **Singapore**: AI Verify Discussion Paper, pages 24-25; Proposed Model Governance Framework for Generative AI, pages 13-14. **South Korea:** Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
123 **Australia**: eSafety Commissioner Position Statement, page 31. **Japan**: Governance Guidelines for Implementation of AI Principles (Ver 1.1), pages 39-43; Guidelines for AI Business Operators, Recommendation P-7. **South Korea:** Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
124 **Australia**: eSafety Commissioner Position Statement, pages 17, 30. **Japan**: Guidelines for AI Business Operators, Recommendation P-7.
125 **Australia**: Safe and Responsible AI in Australia Discussion Paper, pages 40-41; Government's Interim Response to Safe and Responsible AI in Australia Consultation, page 20. **Japan**: Guidelines for AI Business Operators, Recommendation D-6, D-7, P-6. **Singapore**: AI Verify Discussion Paper, page 21.
126 **Australia**: Safe and Responsible AI in Australia Discussion Paper, pages 40-41: eSafety Commissioner Position Statement, pages 7, 23, 31. **China**: Basic Security Requirements, Section 7.
127 **Australia**: Safe and Responsible AI in Australia Discussion Paper, pages 40-41; eSafety Commissioner Position Statement, pages 9, 30. **China**: Interim Generative AI Measures, Article 15. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
128 **Australia**: eSafety Commissioner Position Statement, page 31. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
129 **Australia**: Safe and Responsible AI in Australia Discussion Paper, pages 40-41; eSafety Commissioner Position Statement, pages 9, 30. **China**: Interim Generative AI Measures, Article 15. **South Korea**: Policy Direction for Safe Use of Personal Information in the AI Era, Section IV.
130 **Australia**: eSafety Commissioner Position Statement, page 23. **Singapore**: Model AI Governance Framework, pages 44-45.
131 **Australia**: Government's Interim Response to Safe and Responsible AI in Australia Consultation, pages 13, 20; eSafety Commissioner Position Statement, pages 9, 29-30. **China**: Deep Synthesis Regulations, Articles 16-17, Interim Generative AI Measures, Article 12. **Japan**: Guidelines for AI Business Operators, Appendix 3. **Singapore**: AI Verify Discussion Paper, page 21; Proposed Model Governance Framework for Generative AI, page 17.
132 https://c2pa.org/
133 *Supra* n 109, page 94.
134 *Ibid*; See also Rosenblatt, B (2023). "Google And OpenAI Plan Technology To Track AI-Generated Content" *Forbes.* https://www.forbes.com/sites/billrosenblatt/2023/07/22/google-and-openai-plan-technology-to-track-ai-generated-content/?sh=348a1ece131b
135 **Australia**: Government's Interim Response to Safe and Responsible AI in Australia Consultation, pages 13, 20; eSafety Commissioner Position Statement, pages 9, 29-30. **China**: Deep Synthesis Regulations, Articles 16-17; Interim Generative AI Measures, Article 12. **Singapore**: AI Verify Discussion Paper, page 21; Proposed Model Governance Framework for Generative AI, page 17.
136 **Singapore**: AI Verify Discussion Paper, page 21; Proposed Model Governance Framework for Generative AI, page 17.
137 https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework
138 https://www.chiefscientist.gov.au/sites/default/files/2023-06/Rapid%20Response%20Information%20Report%20-%20Generative%20AI%20v1_1.pdf
139 https://consult.industry.gov.au/supporting-responsible-ai
140 https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/Safe-and-responsible-AI-in-Australia-discussion-paper.pdf

141 https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf
142 https://www.esafety.gov.au/newsroom/media-releases/new-industry-recommendations-to-curb-harms-of-generative-ai
143 https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai
144 https://www.esafety.gov.au/industry/tech-trends-and-challenges
145 https://dp-reg.gov.au
146 https://www.oaic.gov.au/newsroom/digital-platform-regulators-release-working-papers-on-algorithms-and-ai
147 https://dp-reg.gov.au/publications/working-paper-2-examination-technology-large-language-models
148 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/
149 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/
150 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/
151 http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm
152 http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
153 https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf
154 https://www.gov.cn/gongbao/content/2020/content_5492511.htm
155 https://www.gov.cn/zhengce/content/202306/content_6884925.htm
156 http://www.fxcxw.org.cn/dyna/content.php?id=26910
157 https://www8.cao.go.jp/cstp/ai/ningen/ningen.html. English translation available at https://www8.cao.go.jp/cstp/ai/humancentricai.pdf
158 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf
159 https://www.ppc.go.jp/files/pdf/230602_kouhou_houdou.pdf
160 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240119_report.html
161 https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html
162 https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework
163 https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf
164 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
165 https://aiverifyfoundation.sg/what-is-ai-verify/
166 https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox
167 https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf
168 https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf
169 Available at https://openresearch-repository.anu.edu.au/bitstream/1885/277585/1/SKAI_31.pdf. No authoritative English language translation is available.
170 https://eiec.kdi.re.kr/policy/materialView.do?num=208784&topic=&pp=20&datecount=&recommend=&pg=
171 https://www.koreaherald.com/view.php?ud=20201223000794
172 https://likms.assembly.go.kr/bill/billDetail.do?billId=PRC_Y2B1M0R6G2I2P1B0V2X9H4Z0X3M3J2
173 https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&pageIndex=7&bbsSeqNo=94&nttSeqNo=3183387&searchOpt=ALL&searchTxt=
174 https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9055#LINK
175 https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9083
176 https://www.priv.gc.ca/en/opc-news/speeches/2023/s-d_20230621_g7/
177 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document02_en.pdf
178 https://www.oecd.org/publications/g7-hiroshima-process-on-generative-artificial-intelligence-ai-bf3c0c60-en.htm
179 https://www.mofa.go.jp/files/100573471.pdf
180 https://www.mofa.go.jp/files/100573473.pdf
181 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
182 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/.
183 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 Advacne 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/.
184 A comprehensive list of the obligations imposed on agencies by the EO may be found here: https://crsreports.congress.gov/product/pdf/R/R47843
185 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
186 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
187 *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/dl?inline=
188 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-government-related
189 https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

1350 EYE STREET NW | SUITE 350 | WASHINGTON, DC 20005 FPF.ORG | info@fpf.org