The Model AI Governance Framework for Generative AI

Introduction

On 30 May 2024, the Infocomm Media Development Authority (**IMDA**) and the AI Verify Foundation published the Model AI Governance Framework for Generative AI (**Framework**).

The Model AI Governance Framework (**Traditional AI Framework**), which pertained to the development and deployment of traditional artificial intelligence (**AI**) solutions, was first released in 2019, and updated in 2020.[1] According to the IMDA, the Framework expands on the Traditional AI Framework.

The Framework relates to generative AI, which refers to AI models capable of generating text, images or other media types; these models learn the patterns and structure of their input training data and generate new data with similar characteristics.[2] The Framework recognises that generative AI has unique characteristics and involves unique risks. For one, generative AI allows the rapid creation of realistic synthetic content, which makes it harder for consumers to distinguish between AI-generated and original content.

The Framework takes into account the AI risks and policy ideas highlighted in the Discussion Paper on Generative AI: Implications for Trust and Governance, which was issued in June 2023.[3] It also draws from insights and discussions with key jurisdictions, international organisations, research communities and leading AI organisations.[4] The Framework recognises that it will continue to evolve based on engagement with key stakeholders and developments in technology and policy discussions.

The Framework seeks to set out a systematic and balanced approach to address generative AI concerns while continuing to facilitate innovation.[5]

Framework Recommendations

The Framework incorporates various recommendations under nine "dimensions", detailed below.

Accountability

The Framework recommends instituting the right incentive structure for different players in the AI system development life cycle to be responsible to end-users. According to the Framework, such players include model developers, application deployers and cloud service providers, with the latter often providing platforms which AI applications are hosted on.

The Framework specifies that instituting such an incentive structure involves allocating responsibility upfront in the development process (***ex-ante***), and providing guidance on how end-users can obtain redress if issues are discovered after development (**ex-post**).

1 Page 3 of the Framework.
2 Page 3, Footnote 3 of the Framework.
3 Page 3 of the Framework.
4 Page 5 of the Framework.
5 Page 3 of the Framework.

The Framework recommends allocating responsibility *ex-ante* based on each player's level of control in the generative AI development chain, which may in turn vary across model types. Since model developers are the most knowledgeable about their own models and the deployment thereof, the Framework suggests that they lead the development process.

As for ex-post responsibility, the Framework recommends implementing measures relating to indemnity and insurance. One recommended measure is the underwriting of certain risks, including third-party copyright claims. Another is to update legal frameworks to make them more flexible, and to enable them to address emerging risks easily and fairly. The Framework also recommends using solutions such as no-fault insurance to address residual issues, although further study is required regarding insurance in the AI context.

Data

The Framework recognises that data is core to the development of models and applications, because data significantly impacts the quality of the model output. Therefore, the Framework states that it is necessary to undertake data quality control measures, and to create more trusted datasets. For instance, the Framework recommends that Governments curate a repository of representative training datasets for specific contexts.

The Framework also recognises that it is important, in a pragmatic manner, to provide clarity and certainty to businesses on how they can use data in model development, and ensure fair treatment in situations where using data for model training is potentially contentious. Such situations include where the data constitutes publicly available personal data and copyright material. Relatedly, the Framework recognises that it is important to address whether the generation of creative output which may mimic existing creators' styles, amounts to fair use.

Regarding personal data, the Framework recommends that policymakers stipulate how existing personal data laws apply to generative AI. Another recommendation is to explore the use of Privacy Enhancing Technologies to protect data confidentiality and privacy while allowing data to be used to develop AI Models.

Regarding copyright material, the Framework recommends developing approaches to clearly and efficiently resolve issues relating to the use of copyright material in training datasets and fair use. Recommended approaches include non-legislative solutions such as copyright guidelines and codes of practice, and open dialogue among stakeholders.

Trusted development and deployment

The Traditional AI Framework focused on best practices for developing and deploying traditional AI systems; such best practices have been incorporated into and expanded under the Trusted Development and Deployment dimension of the Framework.[6]

The Framework recommends enhancing transparency in relation to baseline safety and hygiene measures which are based on industry best practices in development, evaluation and disclosure. According to the Framework, these measures include using fine-tuning techniques which guide the model to generate safer

[6] Page 3, Footnote 2 of the Framework.

output that is more aligned with human preferences and values, conducting risk assessments, and using user interaction techniques (such as input and output filters) to reduce harmful output.

To enhance transparency, the Framework recommends that model developers standardise the types of information disclosed, such as the data used, training infrastructure, evaluation results, mitigation and safety measures, risks and limitations, intended use, and user data protection. If a model poses potentially high risks (e.g., if the model has implications on national security or society), the Framework recommends that there be greater transparency to the Government.

The recommendations also include adopting a more comprehensive and systematic approach to safety evaluations, and achieving further assurance by defining a baseline set of required safety tests and developing shared resources. In this regard, the Framework recommends that there be coherence between baseline and sector-specific requirements.

However, the Framework caveats that transparency must be balanced with legitimate considerations, including safeguarding business and proprietary information, and preventing bad actors from gaming the system. The Framework suggests that model developers calibrate the level of detail disclosed in order to achieve such a balance.

Incident reporting

The Framework recommends that organisations establish structures and processes to enable incident reporting, which will in turn facilitate timely notification and remediation, and enable AI systems to be continuously improved.

Specifically, the Framework suggests that organisations allow vulnerability reporting before incidents occur, as part of an overall proactive security approach. As for post-incident reporting, the Framework encourages organisations to define severe AI incidents or set the materiality threshold for formal reporting, in order to strike a balance between comprehensive reporting and practicality. Relatedly, the Framework indicates that AI incidents can be reported to the equivalent of Information Sharing and Analysis Centres, which are trusted entities to foster information sharing and good practices, or to relevant authorities, where required by law.

Testing and assurance

According to the Framework, third-party testing and assurance helps to provide external validation and added trust. The Framework recommends developing common standards around AI testing to ensure quality and consistency. Specifically, testing methodologies should be reliable and consistent, and the scope of testing should complement internal testing. The Framework also states that third-party testers must be independent, and recommends the development of an accreditation mechanism to ensure that such testers are independent and competent.

Security

The Framework recognises that generative AI models are threatened by novel threat vectors, which go beyond security risks inherent in any software stack; in other words, they go beyond traditional software security concerns.

The Framework recommends addressing such vectors by adapting security-by-design (which the Framework defines as designing security into every phase of the systems development life cycle) to generative AI's unique characteristics. Such unique characteristics include generative AI's ability to inject natural language, and its probabilistic nature.[7]

Further, the Framework recommends developing new security safeguards to support risk assessment and threat modelling, such as input filters, digital forensics tools, and databases that provide information on adversary tactics, techniques and case studies.

Content provenance

According to the Framework, content provenance entails being transparent to end-users about where content comes from, bearing in mind that AI-generated content can exacerbate misinformation and give rise to potential societal threats, such as undermining the integrity of elections.

Specifically, the Framework recommends implementing technical solutions, such as digital watermarking and cryptographic provenance, to catch up with the speed and scale of AI-generated content. Digital watermarking embeds information within the content so that AI-generated content can be identified, while cryptographic provenance helps to track and verify the digital content origin and any edits made.[8] The Framework also recommends complementing such technical solutions with enforcement mechanisms.

The recommendations also include working with key parties in the content life cycle, such as publishers, to support the embedding and display of digital watermarks and provenance details. Other recommendations include simplifying provenance details to facilitate end-user understanding, and standardising the types of edits to be labelled.

Overall, the Framework encourages organisations to carefully design policies, so that they can be practically used in the correct contexts, bearing in mind that it may not be practically feasible for all content creation, editing or display tools to include the above-mentioned technologies in the near term.

Safety and alignment R&D

The Framework recommends accelerating research and development (**R&D**) through global cooperation among AI Safety Institutes to improve the alignment of models with human intention and values, since today's state-of-the-science regarding model safety does not fully cover all risks. The Framework highlights that such alignment must keep pace with present and future catastrophic risks.

Specifically, the Framework encourages organisations to understand and systematically map the diversity of research directions and methods in the field of safety and alignment, and then apply them in a concerted manner. One area of research (i.e., forward alignment) involves developing more aligned models, while another area (i.e., backward alignment) pertains to evaluating a trained model in order to validate its alignment.[9] The Framework also recommends global cooperation, and identifying and prioritising impactful areas of research.

7 Page 22 of the Framework.
8 Page 24 of the Framework.
9 Page 27 of the Framework.

AI for public good

The Framework recommends harnessing AI to benefit the public by democratising access, improving public sector adoption, upskilling workers, and developing AI systems sustainably.

Specifically, the Framework encourages governments and industry partners to improve awareness and provide support to drive innovation and AI use among small and medium enterprises. The Framework also encourages governments to coordinate resources to support public sector AI adoption, and partner companies and communities on digital literacy initiatives to encourage safe and responsible AI use. Other recommendations include concerted upskilling of the workforce, and the redesigning of jobs.

Further, the Framework indicates that applications involving generative AI should be designed in a human-centric way, in order to yield the intended social and human outcomes. Another recommendation is to ensure the sustainable growth of generative AI, such as by developing suitable technology, tracking and measuring generative AI's carbon footprint, and conducting R&D on green computing techniques.

Conclusion

The Framework is a comprehensive and forward-looking policy document that aims to promote the responsible use of generative AI, while enabling innovation and public good. It provides useful guidance and best practices for key stakeholders in the generative AI ecosystem, including policymakers, industry, researchers and end-users.

Organisations that develop or deploy generative AI solutions need to be aware of the Framework, and take steps to adopt the recommendations set out in the Framework. This may include reviewing and enhancing existing processes and practices, adopting relevant tools and techniques, disclosing relevant information, reporting incidents, conducting third-party testing, and collaborating with other stakeholders.

By doing so, organisations can be accountable to end-users, and contribute to the Framework's broader goal of developing a trusted AI ecosystem, where AI is used for the public good and people can safely and confidently use AI.

If you would like information and/or assistance on the above or any other area of law, you may wish to contact the Partner at WongPartnership whom you normally work with or any of the following Partners:

LAM Chung Nian
Head – Intellectual Property, Technology & Data Group
d: +65 6416 8271
e: chungnian.lam@wongpartnership.com
Click here to view Chung Nian's CV.

Kylie PEH
Partner – Intellectual Property, Technology & Data Group
d: +65 6416 8259
e: kylie.peh@wongpartnership.com
Click here to view Kylie's CV.

Connect with WongPartnership.

WPG MEMBERS AND OFFICES

- contactus@wongpartnership.com

SINGAPORE

-
WongPartnership LLP
12 Marina Boulevard Level 28
Marina Bay Financial Centre Tower 3
Singapore 018982
t +65 6416 8000
f +65 6532 5711/5722

CHINA

-
WongPartnership LLP
Shanghai Representative Office
Unit 1015 Link Square 1
222 Hubin Road
Shanghai 200021, PRC
t +86 21 6340 3131
f +86 21 6340 3315

INDONESIA

-
Makes & Partners Law Firm
Menara Batavia, 7th Floor
Jl. KH. Mas Mansyur Kav. 126
Jakarta 10220, Indonesia
t +62 21 574 7181
f +62 21 574 7180
w makeslaw.com

MALAYSIA

-
Foong & Partners
Advocates & Solicitors
13-1, Menara 1MK, Kompleks 1 Mont' Kiara
No 1 Jalan Kiara, Mont' Kiara
50480 Kuala Lumpur, Malaysia
t +60 3 6419 0822
f +60 3 6419 0823
w foongpartners.com

MIDDLE EAST

-
Al Aidarous Advocates and Legal Consultants
Abdullah Al Mulla Building, Mezzanine Suite 02
39 Hameem Street (side street of Al Murroor Street)
Al Nahyan Camp Area
P.O. Box No. 71284
Abu Dhabi, UAE
t +971 2 6439 222
f +971 2 6349 229
w aidarous.com

-
Al Aidarous Advocates and Legal Consultants
Oberoi Centre, 13th Floor, Marasi Drive, Business Bay
P.O. Box No. 33299
Dubai, UAE
t +971 4 2828 000
f +971 4 2828 011

PHILIPPINES

-
Gruba Law
27/F 88 Corporate Center
141 Valero St., Salcedo Village
Makati City 1227, Philippines
t +63 2 889 6060
f +63 2 889 6066
w grubalaw.com

wongpartnership.com