Model AI Governance Framework for Generative AI

Fostering a Trusted Ecosystem

Published 30 May 2024

AI VERIFY FOUNDATION

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

CONTENTS

EXECUTIVE SUMMARY

Generative AI has captured the world's imagination. While it holds significant transformative potential, it also comes with risks. **Building a trusted ecosystem is therefore critical** — it helps people embrace AI with confidence, gives maximal space for innovation, and serves as a core foundation to harnessing AI for the Public Good.

AI, as a whole, is a technology that has been developing over the years. Prior development and deployment is sometimes termed *traditional AI*.[1] To **lay the groundwork** to promote the responsible use of traditional AI, Singapore released the first version of the *Model AI Governance Framework* in 2019, and updated it subsequently in 2020.[2] The recent advent of *generative AI*[3] has reinforced some of the same AI risks (e.g., bias, misuse, lack of explainability), and introduced new ones (e.g., hallucination, copyright infringement, value alignment). These concerns were highlighted in our earlier *Discussion Paper on Generative AI: Implications for Trust and Governance*,[4] issued in June 2023. The discussions and feedback have been instructive.

Existing governance frameworks need to be reviewed to foster a broader trusted ecosystem. A careful balance needs to be struck between protecting users and driving innovation. There have also been various international discussions pulling in the related and pertinent topics of accountability, copyright and misinformation, among others. These issues are interconnected and need to be viewed in a practical and holistic manner. No single intervention will be a silver bullet.

This **Model AI Governance Framework for Generative AI therefore seeks to set forth a systematic and balanced approach** to address generative AI concerns while continuing to facilitate innovation. It requires all key stakeholders, including policymakers, industry, the research community and the broader public, to collectively do their part. There are nine dimensions which the Framework proposes to be looked at in totality, to foster a trusted ecosystem.

a) **Accountability** — Accountability is a key consideration to incentivise players along the AI development chain to be responsible to end-users. In doing so, we recognise that generative AI, like most software development, involves multiple layers in the tech stack, and hence the allocation of responsibility may not be immediately clear. While generative AI development has unique characteristics, useful parallels can still be drawn with today's cloud and software development stacks, and initial practical steps can be taken.

[1] Traditional AI refers to AI models that make predictions by leveraging insights derived from historical data. Typical traditional AI models include logistic regression, decision trees and conditional random fields. Other terms used to describe this include "discriminative AI".

[2] The focus of the Model AI Governance Framework is to set out best practices for the development and deployment of traditional AI solutions. This has been incorporated into and expanded under the Trusted Development and Deployment dimension of the Model AI Governance Framework for Generative AI.

[3] Generative AI are AI models capable of generating text, images or other media types. They learn the patterns and structure of their input training data and generate new data with similar characteristics. Advances in transformer-based deep neural networks enable generative AI to accept natural language prompts as input, including large language models (**LLM**) such as GPT-4, Gemini, Claude and LLaMA.

[4] The Discussion Paper was jointly published by the Infocomm Media Development Authority of Singapore (**IMDA**), Aicadium and AI Verify Foundation. See https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

b) **Data** — Data is a core element of model development. It significantly impacts the quality of the model output. Hence, what is fed to the model is important and there is a need to ensure data quality, such as through the use of trusted data sources. In cases where the use of data for model training is potentially contentious, such as personal data and copyright material, it is also important to give business clarity, ensure fair treatment, and to do so in a pragmatic way.

c) **Trusted Development and Deployment** — Model development, and the application deployment on top of it, are at the core of AI-driven innovation. Notwithstanding the limited visibility that end-users may have, meaningful transparency around the baseline safety and hygiene measures undertaken is key. This involves industry adopting best practices in development, evaluation, and thereafter "food label"-type transparency and disclosure. This can enhance broader awareness and safety over time.

d) **Incident Reporting** — Even with the most robust development processes and safeguards, no software we use today is completely foolproof. The same applies to AI. Incident reporting is an established practice, and allows for timely notification and remediation. Establishing structures and processes to enable incident monitoring and reporting is therefore key. This also supports continuous improvement of AI systems.

e) **Testing and Assurance** — For a trusted ecosystem, third-party testing and assurance plays a complementary role. We do this today in many domains, such as finance and healthcare, to enable independent verification. Although AI testing is an emerging field, it is valuable for companies to adopt third-party testing and assurance to demonstrate trust with their end-users. It is also important to develop common standards around AI testing to ensure quality and consistency.

f) **Security** — Generative AI introduces the potential for new threat vectors against the models themselves. This goes beyond security risks inherent in any software stack. While this is a nascent area, existing frameworks for information security need to be adapted and new testing tools developed to address these risks.

g) **Content Provenance** — AI-generated content, because of the ease with which it can be created, can exacerbate misinformation. Transparency about where and how content is generated enables end-users to determine how to consume online content in an informed manner. Governments are looking to technical solutions like digital watermarking and cryptographic provenance. These technologies need to be used in the right context.

h) **Safety and Alignment Research & Development (R&D)** — The state-of-the-science today for model safety does not fully cover all risks. Accelerated investment in R&D is required to improve model alignment with human intention and values. Global cooperation among AI safety R&D institutes will be critical to optimise limited resources for maximum impact, and keep pace with commercially driven growth in model capabilities.

i) **AI for Public Good** — Responsible AI goes beyond risk mitigation. It is also about uplifting and empowering our people and businesses to thrive in an AI-enabled future. Democratising AI access, improving public sector AI adoption, upskilling workers and developing AI systems sustainably will support efforts to steer AI towards the Public Good.