PROPOSED MODEL AI GOVERNANCE FRAMEWORK FOR GENERATIVE AI

Fostering a Trusted Ecosystem

Issued 16 January 2024

Recognising the importance of collaboration and crowding in expertise, Singapore set up the AI Verify Foundation to harness the collective power and contributions of the global open-source community to build AI governance testing tools. The mission of the AI Verify Foundation is to foster and coordinate a community of developers to contribute to the development of AI testing frameworks, code base, standards and best practices. It will establish a neutral space for the exchange of ideas and open collaboration, as well as nurture a diverse network of advocates for AI testing and drive broad adoption through education and outreach. The vision is to build a community that will contribute to the broader good of humanity, by enabling trusted development of AI.

At IMDA, we see ourselves as Architects of Singapore's Digital Future. We cover the digital space from end to end, and are unique as a government agency in having three concurrent hats – as Economic Developer (from enterprise digitalisation to funding R&D), as a Regulator building a trusted ecosystem (from data/AI to digital infrastructure), and as a Social Leveller (driving digital inclusion and making sure that no one is left behind). Hence, we look at the governance of AI not in isolation, but at that intersection with the economy and broader society. By bringing the three hats together, we hope to better push boundaries, not only in Singapore, but in Asia and beyond, and make a difference in enabling the safe and trusted use of this emerging and dynamic technology.

TABLE OF 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. The recent advent of *generative AI*[2] 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*,[3] 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, 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] Generative AI are AI models capable of generating text, images or other media. 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.
[3] 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.

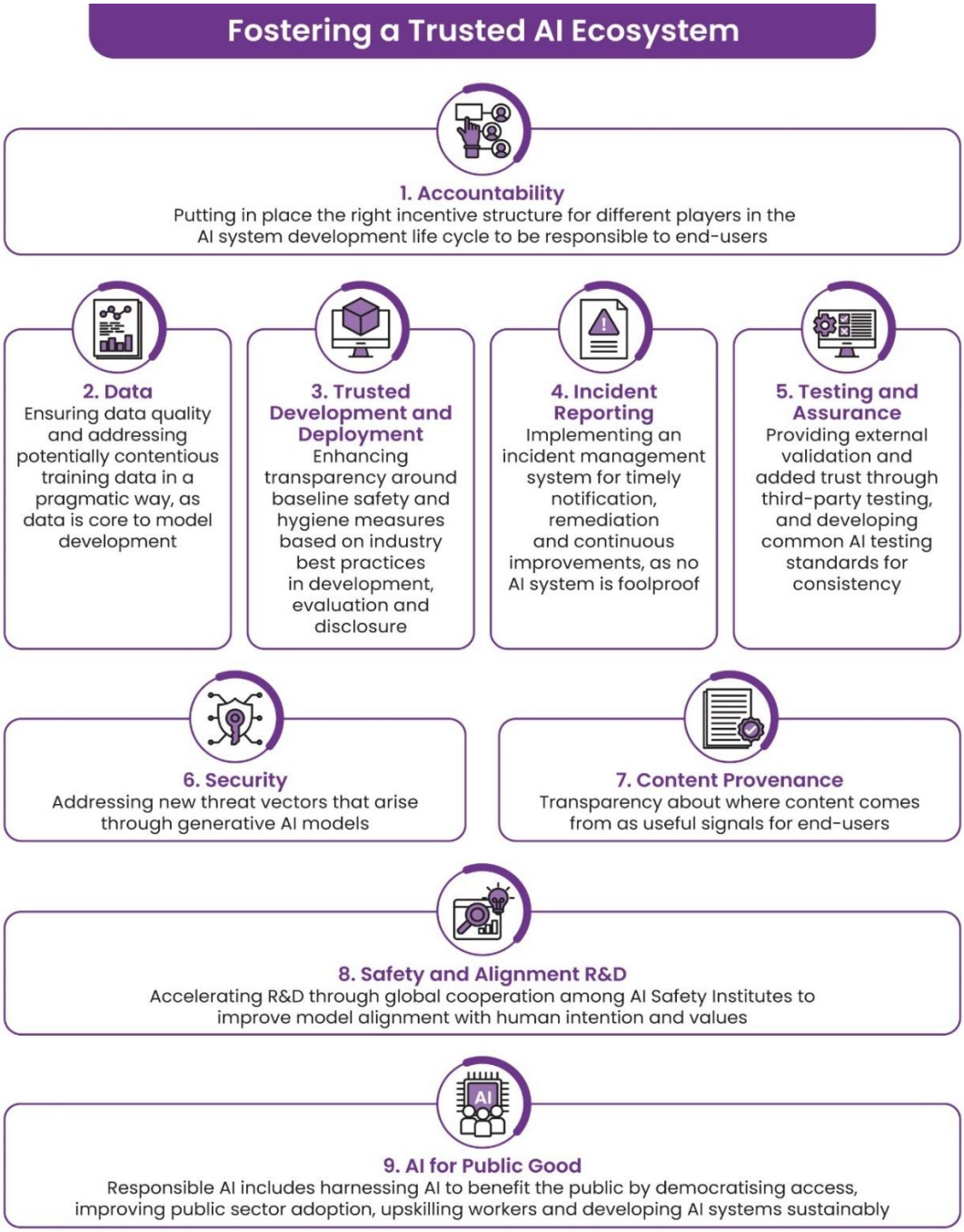
e) **Testing and Assurance** – For a trusted ecosystem, third-party testing and assurance plays a complementary role. We do this today in many domains, such as finance and healthcare, to enable independent verification. Although AI testing is an emerging field, it is valuable for companies to adopt third-party 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 to be injected through 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.



This draft Framework builds on the policy ideas highlighted in our Discussion Paper on Generative AI and draws from insights and discussions with key jurisdictions, international organisations, research communities and leading AI organisations. The recommendations here will evolve as technology and policy discussions develop. We **welcome feedback** to enhance and refine this Model Governance Framework for Generative AI.

1. Accountability

Accountability is a key consideration in fostering a trusted ecosystem. Players along the AI development chain need to be responsible towards end-users, and the structural incentives should align with this need. These players include model developers, application deployers[4] and cloud service providers (who often provide platforms on which AI applications are hosted). Generative AI, like most software development, involves multiple layers in the tech stack. While the allocation of responsibility may not be immediately clear, useful parallels can be drawn with today's cloud and software development, and practical steps can be taken.

Design

To do this comprehensively, there should be consideration for how responsibility is allocated both *upfront* in the development process (ex-ante) as best practice, and guidance on how redress can be obtained if issues are discovered *thereafter* (ex-post).

Ex Ante – Allocation upfront

Responsibility can be allocated based on the level of control that each stakeholder has in the generative AI development chain, so that the able party takes necessary action to protect end-users. As a reference, while there may be various stakeholders in the development chain, the cloud industry[5] has built and codified comprehensive shared responsibility models over time. The objective is to ensure overall security of the cloud environment. These models allocate responsibility by explaining the controls and measures that cloud service providers (who provide the base infrastructure layer) and their customers (who host applications on the layer above) respectively undertake.

There is value in **extending this approach to AI development**. Cloud service providers have recently extended some elements of their cloud shared responsibility models to cover AI, placing initial focus on security controls.[6] This is a good start, and a similar approach can be taken to address other safety concerns. The AI shared responsibility approach may also need to consider different model types (e.g. closed-source, open-source[7] or open-weights[8]), given the different levels of control that application deployers have for each model type. Responsibility in this case, for example when using open-source/weights models, should require application deployers to download models from reputable platforms to minimise the risk of tampered models. Being the most knowledgeable about their own models and how they are deployed**, model developers are well placed to lead this development in a concerted**

[4] We recognise that the generative AI development chain is complex, and that application developers and application deployers can sometimes be two different parties. For simplicity, this paper uses the term "application deployers" to refer to both application developers and deployers.
[5] This includes Google Cloud, Microsoft Azure and Amazon Web Services.
[6] Microsoft, which is both a cloud and model service provider, has initiated some elements of this. See https://learn.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility-ai
[7] Open sourcing makes available the full source code and information required for re-training the model from scratch, including model architecture code, training methodology and hyperparameters, original training dataset and documentation. Models that are closer to this end of the spectrum (but not fully open) include Dolly and BLOOMZ.
[8] Open-weights makes available pre-trained parameters or weights of the model itself, but not the training code, dataset, methodology, etc. Existing open-weights models include LlaMa2, Falcon-40B-Instruct and Mistral 7B-Instruct.

manner. This will provide stakeholders with greater certainty upfront, and foster a safer ecosystem.

Ex Post – Safety Nets

Shared responsibility models serve as an important foundation for accountability – they provide clarity on redress when issues occur. However, **they may not be able to cover all possible scenarios**. Allocating responsibility when there are new or unanticipated issues may also be practically challenging. It will be worth considering additional measures – including concepts around indemnity and insurance – to better cover end-users.

This exists in a limited form today. In clearer areas where redress is needed, the industry has moved. Some model developers[9] have begun to **underwrite certain risks**, such as third-party copyright claims arising from the use of their AI products and services. In doing so, developers implicitly acknowledge their responsibility for model training data and how their models are used.

There will inevitably be other areas that are not as clear and not well-covered. This may include risks that have disproportionate impact on society as a whole, and which may only emerge as AI is used. It is therefore useful to consider **updating legal frameworks** to make them more flexible, and to allow emerging risks to be easily and fairly addressed. This is akin to how end-users of physical products today enjoy safety protections. One example of such efforts is the EU's proposed AI Liability Directive and Revised Product Liability Directive. If adopted, the Directives aim to make it simpler for end-users to prove damage caused by AI-enabled products and services. This ensures that no party is unfairly disadvantaged by the compensation process.

Finally, there are bound to be **residual issues** that fall through the cracks. This is a very nascent discussion, and alternative solutions such as no-fault insurance[10] could be considered as a safety net.

[9] For example, Adobe, Anthropic, Google, Microsoft and OpenAI.
[10] Under a no-fault insurance model, stakeholders' expenses are covered regardless of who is at fault. It is currently adopted in the US for some types of motor accident claims.

2. Data

Data is a core element of model and application development. A large corpus of data is needed to train robust and reliable AI models. Given its importance, **businesses require clarity and certainty on how they can use data in model development.** This includes potentially contentious areas such as publicly-available personal data and copyright material, which are typically included in web-scraped datasets. In such cases, it is important to recognise competing concerns, ensure fair treatment, and to do so in a pragmatic way. In addition, developing a model well requires **good quality** data, and in some circumstances **representative** data as well. It is also important to ensure the **integrity** of available data sets[11].

Design

Trusted use of personal data

As personal data operates within existing legal regimes, a useful starting point is for policymakers to articulate how **existing personal data laws apply to generative AI**. This will facilitate the use of personal data in a manner that still protects the rights of individuals. For example, policymakers and regulators can clarify consent requirements or applicable exceptions, and provide guidance on good business practices for data use in AI.

An emerging group of technologies, known collectively as **Privacy Enhancing Technologies** (**PETs**), has the potential to allow data to be used in the development of AI models while protecting data confidentiality and privacy. Some PETs are not new, such as anonymisation techniques, while other technologies are still nascent and evolving[12]. The understanding of how PETs can be applied to AI will be an important area to advance.

Balancing copyright with data accessibility

From a model development perspective, the use of **copyright material in training datasets** and the issue of consent from copyright owners is starting to raise concerns. Models are also increasingly being used for **generating creative output** – some of which mimic the styles of existing creators and give rise to considerations of whether this would constitute fair use.[13]

[11] Data poisoning attacks training datasets by introducing, modifying, or deleting specific data points. For example, with knowledge of the exact time model developers collect content (e.g. via snapshots) from sources like Wikipedia, bad actors can "poison" the Wikipedia webpages with false content, which will be scraped and used to train the generative AI model. Even if the source moderators undo the changes made to the webpages, the content would have been scraped and used.

[12] IMDA's PETs Sandbox helps to facilitate experimentation based on real-world use cases, including using PETs for AI. This enables industry to explore innovative use of this emerging technology while ensuring PETs are deployed in a safe and compliant manner. See https://www.imda.gov.sg/how-we-can-help/data-innovation/privacy-enhancing-technology-sandboxes

[13] The copyright issue has given rise to varied interests and concerns amongst different stakeholders, with policymakers studying to find the best way forward. Copyright owners have requested remuneration for use of their works to train models, concerned that such systems may compete with them and impact their livelihood. They have advocated licensing-based solutions to facilitate text and data mining activities for machine learning, as well as an opt-out system for copyright owners from statutory exceptions for text and data mining and machine learning activities to avoid unduly impinging on their commercial interests. Others have argued that text and data mining, and machine learning do not infringe copyright because training does not involve the copying and use of the creative expression in works. There are also practical considerations surrounding obtaining consent from every copyright owner, as well as trade-offs in model performance.

Given the large volume of data involved in AI training, there is value in developing approaches to **resolve these difficult issues in a clear and efficient manner**. Today, legal frameworks have not yet coalesced around such an approach. Some copyright owners have instituted lawsuits against generative AI companies in the US and UK courts. Various countries are also exploring non-legislative solutions such as copyright guidelines[14] and codes of practice for developers and end-users[15].

Given the various interests at stake, **policymakers should foster open dialogue** amongst all relevant stakeholders, to understand the impact of the fast-evolving generative AI technology, and ensure that potential solutions are balanced and in line with market realities.

Facilitating access to quality data

As an overall hygiene measure at an organisational level, it would be good discipline for AI developers to **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.

Globally, it is worth considering a concerted effort to **expand the available pool of trusted data sets**. Reference data sets are important tools in both AI model development (e.g. for finetuning) as well as benchmarking and evaluation.[16] Governments can also consider working with their local communities to **curate a repository of representative training data sets for their specific context** (e.g. in low resource languages). This helps to improve the availability of quality datasets that reflect the cultural and social diversity of a country, and in turn supports the development of safer and more culturally representative models.

[14] Japan and the Republic of Korea have announced the development of copyright guidelines to address generative AI issues, though they have not yet been issued.

[15] UK has announced that it is developing a voluntary code of practice between end-users and rights holders through a working group with diverse participation from technology, creative and research sectors. The stated aims of the working group are to make licenses for data mining more available, to help to overcome barriers that AI firms and end-users currently face, and to ensure there are protections for rights holders.

[16] This is akin to reference standards in, for example, the pharmaceutical industry, which are used as a basis for evaluation for drugs.

3. Trusted Development and Deployment

Model development, and the application deployment on top of it, are at the core of AI-driven innovation. Today, however, there is a lack of information on the approaches being taken to ensure trustworthy models. Even in cases of "open-source" models, some important information like the methodology and datasets may not be made available.

Going forward, it is important that the industry coalesces around best practices in development and in turn safety evaluation. Thereafter, meaningful transparency around baseline safety and hygiene measures undertaken will also be key. This will need to be balanced with legitimate considerations such as safeguarding business and proprietary information, and not allowing bad actors to game the system.

Design

Safety best practices need to be implemented by model developers and application deployers across the AI development lifecycle, around **development, disclosure and evaluation.**

Development – Baseline Safety Practices

Safety measures are developing rapidly and model developers/application deployers are best placed to determine what to use. Even so, **industry practices are starting to coalesce around some common safety practices**.

For example, after pre-training, fine-tuning techniques such as Reinforcement Learning from Human Feedback (**RLHF**)[17] can guide the model to generate safer output that is more aligned with human preferences and values. A crucial step for safety is also to consider the context of the use case and conduct a risk assessment. For example, further fine-tuning or using user interaction techniques (such as input and output filters) can help to reduce harmful output. Techniques like Retrieval-Augmented Generation (**RAG**)[18] and few-shot learning are also commonly used to reduce hallucinations and improve accuracy.

Disclosure – "Food Labels"

Transparency around these safety measures undertaken, that form the core of the AI model's make-up, is then key. This is **akin to "food/ingredient labels"**. By providing relevant information to downstream users, they can make more informed decisions. While leading model developers already disclose some information, **standardising disclosure** will facilitate comparability across models and promote safer model use. Relevant areas may include:

a) Data used**:** An overview of the types of training data sources and how data was processed before training.

[17] RLHF is a technique used to improve LLMs by using human feedback to train a preference model, that in turns trains the LLM using reinforcement learning.

[18] RAG is a technique that helps models provide more contextually appropriate and current responses that are specific to an organisation or industry. This is done by linking generative AI services to external resources, thereby giving models sources to cite and enhancing the accuracy and reliability of generative AI models with facts fetched from trusted sources.

b) Training infrastructure**:** An overview of the training infrastructure used and, where possible, estimated environmental impact[19].

c) Evaluation results**:** Overview of evaluations done and key results.

d) Mitigations and safety measures**:** Safety measures implemented (e.g. bias correction techniques).

e) Risks and limitations**:** Model's known risks and moves to address these risks.

f) Intended use**:** Clear statement setting out the scope of the model's intended use.

g) User data protection**:** Outlining how users' data will be used and protected.

The level of detail disclosed can be calibrated based on the need to be transparent vis-à-vis protecting proprietary information. One step forward would be for the industry to agree on the baseline transparency to be provided as part of general disclosure to all parties. This involves both the model developers and application deployers. Alternatively, the development of such a baseline can be facilitated by governments and third parties.

Greater transparency to government will also be needed for models that pose potentially high risks, such as advanced models that have national security or societal implications. There is therefore space for policymakers to define the model risk thresholds, above which additional oversight measures would apply.

Evaluation

There are generally two main approaches to evaluate generative AI today – (i) **benchmarking** tests models against datasets of questions/answers to assess performance and safety; and (ii) **red teaming** where a red team acts as an adversarial user to "break" the model and induce safety, security and other violations. Although benchmarking and red teaming are commonly adopted today, they still fall far short in terms of giving a robust assessment of model performance and safety (see the section on Safety and Alignment R&D).

Even within the benchmarking and red teaming framework, most evaluation today focuses on generative AI's front-end performance, and less about its back-end safety. There is also a lack of evaluation tools (e.g. for multi-modal models), as well as testing for dangerous capabilities. Another issue is in consistency – many tests and evaluations today need to be customised to a specific model and at times, comparability is a challenge.

There is therefore a need to **work towards a more comprehensive and systematic approach to safety evaluations**. This will yield more useful and comparable insights. To provide additional assurance, the standardised approach could also include defining a baseline set of required safety tests, in consultation with policymakers.

[19] More so as AI training and the use of accelerated compute is driving up carbon emissions.

A Starting Point for Standardised Safety Evaluations

AI Verify Foundation and IMDA recommended an **initial set of standardised model safety evaluations** for LLMs**,** covering robustness, factuality, propensity to bias, toxicity generation and data governance. It can be found in the paper titled *Cataloguing LLM Evaluations* issued in October 2023.[20] The paper provides both a landscape scan as well as practical guidance on what safety evaluations may be considered. These recommendations have to be continuously improved, given rapid advances in the generative AI space.

Sectors and domains may have unique needs that require additional evaluations (e.g. mandating stringent accuracy thresholds for high-risk use cases such as medical diagnosis). Application deployers, additionally, will more likely focus on domain-specific assessments that address their use cases. Industry and sectoral policymakers need to jointly improve evaluation benchmarks and tools, while still **maintaining coherence between baseline and sector-specific requirements**.[21]

[20] See aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf
[21] For example, aligning safety principles, using common terminologies.

4. Incident Reporting

Even with the most robust development processes and safeguards, no software that we use today is foolproof. The same applies to AI. Incident reporting is an established practice, including in critical domains such as telecommunications, finance and cybersecurity. It allows for timely notification and remediation. **Establishing the structures and processes to enable incident reporting** is therefore key. This in turn supports continuous improvement of AI systems through insights, remediation and patching.

Design

Vulnerability Reporting – Incentive to Act Pre-Emptively

Before incidents happen, software product owners adopt vulnerability reporting as part of an overall proactive security approach. They co-opt/support white hats or independent researchers to discover vulnerabilities in their software, sometimes through a curated bug-bounty programme. Once discovered, a vulnerability is reported and the product owner is then given time (typically 90 days based on industry practice) to patch their software, publish the vulnerability (such as by filing a CVE – see box below) and crediting the white hat/independent researcher. This allows both the software product owner and users to undertake proactive steps to enhance overall security.

Common Vulnerabilities and Exposures (CVE) Programme

The CVE programme, managed by the MITRE Corporation, compiles a list of publicly known security vulnerabilities and exposures. This list is widely referred to by cybersecurity teams around the world to look for new vulnerabilities that might affect one's organisation. Software product owners may file vulnerabilities as a CVE. The ability to discover zero-day CVEs is also viewed as an achievement among the white hat community.

AI developers can apply this similar concept, by allowing **reporting channels** for uncovered safety vulnerabilities in their AI systems. They can apply the same best practices for vulnerability reporting, including a time-window to assess the incident, patch and publish.

Incident Reporting

After incidents happen, organisations need internal processes to report the incident for timely notification and remediation. Depending on the impact of the incident and how extensively AI was involved, this could include notifying both the public as well as governments. Defining "severe AI incidents" or setting the materiality threshold for formal reporting is therefore key. Borrowing from cybersecurity, 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, as well as to relevant authorities where required by law.

Reporting should be proportionate, which means striking a balance between comprehensive reporting and practicality. This will need to be calibrated to suit the specific local context. In this regard, the impending *EU AI Act* provides one reference point for legal reporting requirements (see box below).

Incident Reporting Under the Impending EU AI Act

Providers of high-risk AI systems are required to report serious incidents to the market surveillance authorities of the Member States where that incident occurred, within 15 days after the AI system provider becomes aware of the incident. "**Serious incident**" is defined as any incident or malfunctioning of an AI system that directly or indirectly leads to the death of a person, serious damage to a person's health, serious and irreversible disruption of critical infrastructure, breaches of fundamental rights under Union law, or serious damage to property or the environment.

5. Testing and Assurance

Third-party testing and assurance often plays a complementary role in a trusted ecosystem. We do this today in many domains, such as finance and healthcare, to enable independent verification. While companies typically conduct audits to demonstrate compliance with regulation, more companies are beginning to see external audits as a **useful mechanism to provide transparency and build greater credibility and trust with end-users**[22].

While this is an emerging field, we can **draw from established audit practices** to grow the AI third-party testing ecosystem. Third-party testing will also benefit from comprehensive and consistent standards around AI evaluations (discussed earlier in the section on Trusted Development and Deployment).

Design

Fostering development of a third-party testing ecosystem involves two pivotal aspects:

a) **How to test**: Defining a testing methodology that is reliable and consistent.

b) **Who to test**: Identifying the entities to conduct testing that ensures independence.

How to test - Standardisation

In the near term, third-party testing will comprise the same set of benchmarks and evaluation used by developers themselves[23]. Eventually, this needs to be done in a **standardised way** for third-party testing to be effective, and to facilitate meaningful comparability across models.

Greater emphasis should therefore be placed on **setting common benchmarks and methodologies**. This may be catalysed by having common tooling to reduce the friction required to test across different models or applications. Thereafter, for more mature areas, AI testing could be codified through standards organisations like ISO/IEC and IEEE, to support more harmonised and robust third-party testing.

Who to test – Trusted Accreditation

Independence is key to ensuring the objectivity and integrity of test results. Building up a pool of qualified third-party testers is critical. Concerted efforts by industry bodies and governments will be useful to grow capabilities in this area. Eventually, an **accreditation mechanism** could be developed to ensure independence and competency. This is common practice in many domains (e.g. finance). Many audit and professional services firms are understandably increasingly keen to grow some initial AI audit capability and services.

[22] For instance, in the White House Voluntary Commitments, several AI companies pledged to conduct external model red teaming as a means of demonstrating trust.
[23] Stanford's Holistic Evaluation of Language Models is an example of a third-party conducting benchmark tests today.

6. Security

Generative AI has brought renewed focus on the security of AI itself. Many issues are familiar, such as supply chain risks in AI/ML middleware. Others are distinct to generative AI, such as prompt attacks injected through the model architecture, which allows attackers to, for example, exfiltrate sensitive information/model weights. In addressing AI security, it is useful to separate **traditional software security concerns** addressed via current approaches, from **novel threat vectors against the AI model itself.** The latter is a nascent space. Nevertheless, similar security concepts may still apply.

Design

Adapt "Security-by-Design"

Security-by-design is a fundamental security concept. It seeks to minimise system vulnerabilities and reduce the attack surface through designing security into every phase of the systems development lifecycle (**SDLC**). Key SDLC stages include development, evaluation, operations and maintenance.

However, **refinements may be needed given the unique characteristics of generative AI**. For example, the ability to **inject natural language** as input can pose challenges in designing appropriate security controls[24]. Furthermore, the **probabilistic nature** of generative AI challenges traditional evaluation techniques that inform system refinement and risk mitigation in the SDLC. Hence, new concepts have to be developed/adapted for generative AI.

Develop New Security Safeguards

New tools have to be developed and may include:

a) **Input Filters:** Input moderation tools detect unsafe prompts (e.g. blocking malicious code). The tools need to be tailored to understand domain-specific risks.

b) **Digital Forensics Tools for Generative AI:** Digital forensics tools are used to investigate and analyse digital data (e.g. file contents) to reconstruct a cybersecurity incident. Existing forensics tools should be improved with new techniques to identify and extract malicious codes that might be hidden within a generative AI model.

Apart from these tools, databases such as MITRE's Adversarial Threat Landscape for AI Systems provide information on adversary tactics, techniques and case studies for machine learning systems, including generative AI. AI developers can use these to support risk assessment and threat modelling, and to identify useful tools/processes.

[24] This is because existing security controls, such as next-generation firewalls and data loss protection typically rely on restricting communication protocols between nodes and establishing pre-defined filters to detect and mitigate malicious attacks. They therefore do not perform well with wide-ranging communications that may span interactive and dynamic dialogue, long text and source code. In the case of multi-modal models, this can even extend to various forms of content such as images, videos and sounds.

7. Content Provenance

The rise of generative AI, which enables the **rapid creation of realistic synthetic content**[25] **at scale**, has made it harder for consumers to distinguish between AI-generated and original content. A common manifestation of such concern is deepfakes. This has exacerbated harms like misinformation, and even potential societal threats like undermining the integrity of elections.

There is recognition across governments, industry and society on the **need for technical solutions**, such as digital watermarking and cryptographic provenance, to catch up with the speed and scale of AI-generated content[26]. Digital watermarking and cryptographic provenance both aim to label and provide additional information, and are used to flag content created with or modified by AI.

Digital watermarking techniques embed information within the content and can be used to identify AI-generated content. There are several digital watermarking solutions to label AI-generated content today (e.g. Google DeepMind's SynthID and Meta's Stable Signature). However, it is only possible to decode a watermark through the same company that encodes the watermark[27], due to the current lack of interoperable standards.

Cryptographic provenance solutions track and verify the digital content origin and any edits made, with the records cryptographically protected. The Coalition for Content Provenance and Authenticity (**C2PA**)[28] is driving development of an open standard to enable the tracking of content provenance.

Design

Policies need to be carefully designed to enable practical use in the right contexts. Practically, it may not be feasible for all content creation, editing or display tools to include these technologies in the near term. Provenance information can also be stripped[29]. In addition, consumer understanding of these tools is low. Malicious actors will also find ways to circumvent these tools, or worse, use them to create a false sense of authenticity.

There is therefore a need to work with **key parties in the content lifecycle**, such as working with publishers to support the embedding and display of digital watermarks and provenance details. As most digital content is consumed through social media platforms, browsers, or media outlets, publishers' support is critical to provide end-users with the ability to verify

[25] Image, video or audio.
[26] For example, China's *Deep Synthesis Regulations* require watermarking of AI-generated content, the US *Executive Order on the Safe, Secure and Trustworthy Development and Use of AI* commits the government to the development of effective labelling and content provenance mechanisms, and the impending *EU AI Act* imposes specific transparency obligations for deepfake systems.
[27] In the encoding process, a content creator inserts the invisible watermark via an algorithm into the digital image. For decoding, the image is scanned via an algorithm for the presence of an embedded watermark.
[28] This is driven by several companies, including Adobe and Microsoft.
[29] For example, removed by online tools or when uploaded on some online platforms.

content authenticity across various channels. There is also a need to ensure proper and secure implementation, to circumvent bad actors trying to exploit it in any way.

Different types of edits (e.g. whether an image is entirely AI-generated or only a small portion of it is) will impact how the content is perceived by the end-user. To improve end-user experience and enable consumers to discern between non-AI and AI-generated content, **standardising the types of edits to be labelled** would be helpful.

End-users need **greater understanding of content provenance** across the content lifecycle and to learn to utilise tools to verify for authenticity. Key stakeholders (e.g. content creators, publishers, solution providers) can partner policymakers to raise awareness. Provenance details to be displayed should also be simplified to the extent possible to facilitate end-user understanding.

8. Safety and Alignment R&D

Safety techniques, and evaluation tools today do not fully address all potential risks. For example, even RLHF, the primary method for value alignment today, has limitations. Existing large models also lack interpretability and may not be consistently reproducible. Given the speed of model advancement, there is a **need to ensure that human capacity to align and control generative AI keeps pace** with the potential risks, including catastrophic risks**.**

Design

While the call to invest more in R&D is a no-regrets move, there may be practical steps to enhance the speed of translation and use of new R&D insights. There is a need to, for example, **understand and systematically map** the diversity of research directions and methods that have emerged in safety and alignment – and apply them in a concerted manner.

a) One broad area of research entails the **development of more aligned** models (also known by some as "forward alignment")[30], such as through Reinforcement Learning from AI Feedback (**RLAIF**)[31]. RLAIF seeks to improve on RLHF by enhancing feedback efficiency and quality, and enabling scalable oversight of advanced models. It also, however, comes with its own drawbacks.

b) Another area of research is the **evaluation of a model after it is trained, to validate its alignment** (also known by some as "backward alignment"). This includes testing for emergent capabilities so that potentially dangerous abilities, such as autonomous replication and long horizon planning, can be detected early. Mechanistic interpretability, which seeks to understand the neural networks of a model to find the source of problematic behaviours, is also gaining traction as a research area.

To keep pace with advancements in model capabilities, **R&D in model safety and alignment needs to be accelerated**. Today, the majority of alignment research is conducted by AI companies. The setting up of **AI safety R&D institutes or equivalents in UK, US, Japan and Singapore**[32] is therefore a positive development signalling commitment to invest additional resources to drive research for the global good**.**

However, **global cooperation** will be critical to optimise limited talent and resources for maximum impact. Impactful areas of research can be collectively identified and prioritised based on the landscape map. The goal is to enable more impactful R&D efforts to develop safety and evaluation mechanisms ahead of time.

[30] A November 2023 paper on the overview of safety and alignment research termed "forward alignment" and 'backward alignment' as the two key categories of research in this field (Ji et al., 2023, "AI Alignment: A Comprehensive Survey") https://doi.org/10.48550/arXiv.2310.19852

[31] RLAIF uses AI to generate feedback to train the preference model, based on parameters defined by humans. Anthropic's Constitutional AI is an example of RLAIF.

[32] **Singapore's Digital Trust Centre** (**DTC**) looks at overall Digital Trust, including Trusted AI R&D. The DTC is funded by a S$50 million initial investment from IMDA and the National Research Foundation, and was set up in June 2022 to lead Singapore's research and development efforts for trustworthy AI technologies and other trust technologies.

9. AI for Public Good

The transformative potential of generative AI is powerful. If we get the approach correct, global communities will reap exponential benefits. The imperative is to **turbocharge growth** and productivity for developed and developing countries alike, while **empowering people and businesses** globally, because the power of AI is potentially democratising. In this regard, countries must come together to support each other, especially through international and regional groupings. Beyond the large and developed countries (e.g. through G7), this is especially pertinent for developing countries and small states, through key platforms like the Digital Forum of Small States (**Digital FOSS**) at the United Nations, and the Association of Southeast Asian Nations (**ASEAN**). The aim is to establish a global Digital Commons – a place with common rules-of-the-road and equal opportunities for all citizens to flourish, regardless of their geographical location.

Design

There are **four concrete touchpoints** where AI can have beneficial and long-term effects.

Democratising Access to Technology

All members of society should have access to generative AI, done in a trusted manner. Generative AI is inherently intuitive given the natural language focus, but it is still important that the overall product (of which generative AI is just one component) is **designed in a human-centric way**. Most citizens of the world may not understand the technology and the "black-box" underpinning the application they are using. Therefore, designing applications to elicit the intended social and human outcomes is key.

To more broadly support this, **governments can partner companies and communities on digital literacy initiatives** to encourage safe and responsible AI use. Topics could include educating end-users on how to use chatbots safely, sensitising them against "anthropomorphising" AI, and identifying deepfakes.

The adoption of generative AI can also be challenging, especially for small and medium enterprises (**SMEs**). Governments and industry partners can **improve awareness and provide support to drive innovation and AI use among SMEs**. An example is Singapore's Generative AI Sandbox, which provides SMEs with tools and training on generative AI enterprise solutions.[33]

Public Service Delivery

AI should serve the public in impactful ways. Today, AI powers many public services, such as adaptive learning systems in schools and health management systems in hospitals. This unlocks new value propositions, creates efficiencies and improves user experience.

[33] See https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

It is desirable for governments to coordinate resources to support public sector AI adoption. This includes facilitating data sharing across different government agencies, access to high performance compute and other related policies. AI developers play a contributing role by helping governments identify use cases and providing AI solutions to address citizen pain points.

Workforce

For the productive value of AI to be unlocked, concerted **upskilling of the workforce is important.** This is key to countering the potentially negative outcomes of technology replacing labour. Beyond the specific skill sets in using AI tools, other core skills such as creativity, critical thinking and complex problem-solving, are important to helping people harness AI effectively.

Industry, governments and educational institutions can **work together to redesign jobs and provide upskilling opportunities** for workers. As organisations adopt enterprise generative AI solutions, they can also develop dedicated training programmes for their employees. This will enable them to navigate the transitions in their jobs and enjoy the benefits which result from job transformations.

Sustainability

Sustainable growth is key. The power requirements of generative AI hardware are non-trivial and will likely impact sustainability goals. Stakeholders in the generative AI ecosystem therefore need to work together to develop suitable technology (e.g. energy efficient compute) in support of our climate responsibilities.

To inform such plans, the carbon footprint of generative AI will also need to be tracked and measured. AI developers and equipment manufacturers are better placed to conduct R&D on green computing techniques and adopt energy-efficient hardware. In addition, AI workloads can be hosted in data centres that drive best-in-class energy efficiency practices, with green energy sources or pathways.

CONCLUSION

As generative AI continues to develop and evolve, there is a need for **global collaboration on policy approaches**. The nine dimensions in this Framework provide a basis for global conversation to address generative AI concerns while maximising space for continued innovation. The ideas proposed seek to also further the core principles of accountability, transparency, fairness, robustness and security. They reiterate the need for policymakers to work with industry, researchers and like-minded jurisdictions. We hope that this serves as a next step towards developing a **trusted AI ecosystem**, where AI is harnessed for the Public Good, and people embrace AI safely and confidently.

SUBMISSION OF COMMENTS

We welcome feedback and input, to help refine the proposed Framework. All submissions should aim to be concisely written and provide a reasoned explanation. Where feasible, please identify the specific section on which the comments are made.

Comments should be emailed to info@aiverify.sg, with the email header: “Comments on the Proposed Model Governance Framework for Generative AI”.

All submissions should reach us by 15 March 2024.