The Dialogue™
INFORM ENGAGE IDEATE

Written Comments

The Dialogue's comments to Singapore's Proposed Model Governance Framework for Generative AI

March 15, 2024

Written Comments

The Dialogue’s comments to Singapore’s Proposed Model Governance Framework for Generative AI

Authors - *Bhoomika Agarwal, Jameela Sahiba and Pranav Bhaskar Tiwari*

Designer - *Shivam Kulshrestha*

The Dialogue™ is a public policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, on areas around technology and development issues. The DialogueTM has been ranked as the world's Top 10 think tanks to watch out for, by the Think Tank and Civil Societies Programme (TTCSP), University of Pennsylvania in their 2020 and 2021 rankings.

For more information
https://thedialogue.co

Suggested Citation
Agarwal, B., Sahiba, J. & Tiwari, P. B. (March, 2024) Written Comments: The Dialogue's comments to Singapore's Proposed Model Governance Framework for Generative AI. The Dialogue

Catalogue No
TD/ET/WC/0324/03

Publication Date
March 20, 2024

Disclaimer
The facts and information in this report may be reproduced only after giving due attribution to the authors and The Dialogue™.

Singapore recently released a Model AI Governance Framework for Generative AI with the intent to *address generative AI concerns while continuing to facilitate innovation.* The proposed framework builds upon nine parameters to facilitate a trusted ecosystem. The Singapore government has opened a public consultation on its draft AI framework.

The Dialogue, a leading technology policy think tank in India with over 7 years of experience, has been actively involved in research and policy discourse at the intersection of technology, law and society. Our extensive research includes research around artificial intelligence and we have so far published two comprehensive papers on AI: "Principles for Enabling Responsible AI Growth in India: An Ecosystem-Level Approach" and "Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption." These papers underscore the imperative of approaching AI regulation from a principle-level standpoint, delineating key principles essential for fostering responsible and trustworthy AI development and adoption. Notably, our research resonates with the principles outlined in Singapore’s Draft Framework, particularly around transparency, data privacy, accountability, and safety. Through this public consultation, we endeavor to provide a comprehensive response to the draft Framework, meticulously addressing each of the nine principles delineated.

S.No	Aspect	Recommendations	Response
1.	**Data**	• Clarity should be provided on how existing data protection laws apply to generative AI. • Balancing copyright with data accessibility: Need for more dialogue around how to address copyright concerns in training datasets. • Developers should undertake data quality control measures, and adopt general best practices in data governance.	We agree with the recommendation on the need for clarity on the extent of application of existing data frameworks to generative AI. On balancing *copyright with data accessibility,* we submit that the application and training of generative AI within the bounds of copyright law can be aptly described under the fair use exception, employing a method known as non-expressive copying. Non-expressive copying refers to the use of data not for its original, creative content but for its functional value in training AI models. This approach diverges significantly from traditional forms of replication, which often focus on reproducing the expressive elements of copyrighted material. Legal precedents such as Sega v. Accolade[1] and Kelly v. Arriba[2] have ruled on the legitimacy of this transformative approach, highlighting its crucial role in the advancement of technology. Despite the extensive use of data by generative AI, these models are distinct in that they do not store specific content. Instead, they assimilate and learn from overarching patterns, setting them apart from direct forms of copying. While significant, the potential market impact of generative AI is generally consistent with the principles of fair use. This is exemplified in landmark cases like Google v. Oracle,[3] where the Court found that while Google did copy Java API code, this act differed from traditional copying as it was

[1] Sega Enterprises Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992)

[2] Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).

[3] Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183 (2021). See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258 (2023).

S.No	Aspect	Recommendations	Response
			aimed at creating a new, transformative platform, and not for expressing the same content, thus qualifying as a fair use exception in copyright law. As generative AI continues to evolve, it becomes increasingly important for models to cite their sources, particularly when relying on copyrighted data from web sources. This practice enhances transparency and aligns with ethical standards, ensuring respect for the original creators and their economic rights. By balancing innovation with intellectual property protection, this approach promotes a progressive and equitable landscape in the realm of copyright law and AI development. We further agree with the suggestion on the need for developers to adopt adequate data quality safeguards. We suggest following methods that can be adopted towards the same: -Secure Data Sharing Protocols: Implement secure APIs and data sharing protocols to ensure that data is encrypted during transmission. -Transparent Data Usage Policies: Deployers should provide clear data usage policies to their customers. -Data Minimization: Developers can reduce the risk of data breaches by only collecting and using the minimum amount of data required for AI models. -Risk Management Strategies: Develop documented risk management strategies focused on mitigating risks related to data quality and algorithm

S.No	Aspect	Recommendations	Response
			vulnerabilities in response to regulatory changes. Documented risk management strategies refer to well-defined plans and protocols that outline how an organization intends to identify, assess, and address risks associated with specific aspects of its operations. These strategies involve creating clear documentation that articulates the steps and measures to be taken to ensure data quality, address algorithm vulnerabilities, and adapt to regulatory shifts. The documentation may include detailed risk assessment procedures, preventive measures, and response protocols to minimize the impact of potential threats, providing a systematic and organized approach to risk management in the specified domains. -Adequate safeguards: The developer should put safeguards to prevent re-identification from datasets and data leakages. For instance, developers can use techniques to anonymise data. Anonymizing and de-identifying data involve removing or encrypting personally identifiable information (PII) from datasets. PII includes information such as names, addresses, and social security numbers. By using anonymized data, developers can reduce the risk of exposing sensitive information and still derive valuable insights from the data.

S.No	Aspect	Recommendations	Response
2.	**Accountability**	• The framework proposes allocating responsibility on the basis of the level of control that each stakeholder has in the generative AI development chain. • Existing legal frameworks may need adjustments to address new risks emerging from AI use. • Establishing safety nets: No-fault insurance might be considered as a safety net for situations outside the legal framework.	We agree that existing legal frameworks should be updated to accommodate concerns around generative AI. We also welcome the suggestion that all stakeholders be involved in allocation of liability. In our paper *"Towards Trustworthy AI-Sectoral Guidelines for Responsible Adoption"*[4], we identify 'Accountability' as a critical principle that underpins the entire lifecycle of an AI system.[5] It demands that all stakeholders involved in the development and deployment of AI systems take responsibility for ensuring that the technology aligns with human values. However, it might be difficult to quantify the level of control for each stakeholder in the supply chain. We propose that accountability is achieved through careful product design, reliable technical architecture, and a thorough assessment of potential impacts.
3.	**Trusted Development and Deployment**	• Safety best practices need to be implemented by model developers and application deployers across the AI development lifecycle, around development, disclosure and evaluation. • Establishing industry-wide agreements on baseline transparency for model developers and deployers.	We agree with the suggestion to adopt baseline safety practices. However, it would be crucial to note that different AI systems would require different levels of safety standards that would be proportional to the risk of potential harm that might occur in case of an unsafe system. Therefore, developers should undertake requisite action to implement a set of safety assessment standards that would be unique to each system. Technical tools such as risk matrices, failure mode and effects analysis (FMEA), or probabilistic risk assessment (PRA) can be utilized to systematically analyze

[4] Vedashree, R., Sahiba, J., Agarwal, B. & Shekar, K.(2024, February). Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption. The Dialogue https://thedialogue.co/publication/research-paper-towards-trustworthy-ai-sectoral-guidelines-for-responsible-adoption/

[5] Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. https://doi.org/10.1007/s00146-023-01635-y

S.No	Aspect	Recommendations	Response
			and quantify potential risks.[6] We also propose adoption of regulatory sandboxes to pre-assess the impact of safety practices. Regulatory sandboxes offer developers and users the opportunity to test AI systems in a live setting, assessing their robustness and identifying potential concerns. At the development level, developers should actively seek certification and accreditation mechanisms to demonstrate the reliability and robustness of their AI systems. Certifications such as ISO standards for AI[7] can establish adherence to globally recognized best practices, serving as a benchmark for excellence. Accreditation from reputable institutions or industry-specific bodies adds credibility, providing tangible assurances of the system's robustness. By proactively pursuing these mechanisms, developers not only showcase their commitment to quality but also contribute to building trust and confidence among users and stakeholders. At the deployment stage, we propose consistent monitoring and updation of AI models. Towards this, regular surveys and assessments of developments can play a crucial role in staying up-to-date with the latest advancements and breakthroughs in the field. Deployers should identify emerging trends, novel treatment methods, and changes in best practices and ensure timely updation of AI models in line with

[6] Qin, J., Yan, X., & Pedrycz, W. (2020). Failure mode and effects analysis (FMEA) for risk assessment based on interval type-2 fuzzy evidential reasoning method. Applied Soft Computing, 89, 106134. https://doi.org/10.1016/j.asoc.2020.106134
[7] ISO. (2023, September 21). Artificial intelligence (AI) standards. https://www.iso.org/sectors/it-technologies/ai

S.No	Aspect	Recommendations	Response
			these developments. Additionally, emergency shutdown protocols, akin to "kill switches," for high risk AI systems can be implemented. These protocols serve as a safety net, allowing the immediate shutdown of an AI-based system in high-risk circumstances.
4.	**Incident Reporting**	• Reporting can be public or to governments depending on severity, and there should be a balance between comprehensiveness and practicality when setting up such a system.	The recommendation reflects a thoughtful approach to incident reporting for generative AI systems, recognizing the need for transparency, accountability, and practicality in addressing potential issues that may arise in the deployment and operation of such systems. While it's crucial to gather comprehensive data to understand the scope and nature of incidents, it's equally important to ensure that the reporting process remains practical and feasible for stakeholders involved. Overly burdensome reporting requirements could deter participation and hinder the effectiveness of the system. Additionally, deployers may be required to develop and regularly update an incident response plan to effectively respond to and mitigate security incidents. This should include procedures for identifying, reporting, and responding to security breaches.
5.	**Testing and Assurance**	Fostering development of a third-party testing ecosystem: 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.	Developers should actively seek certification and accreditation mechanisms to demonstrate the reliability and robustness of their AI systems. Certifications such as ISO standards for AI[8] can establish adherence to globally recognized best practices,

[8] ISO. (2023, September 21). Artificial intelligence (AI) standards. https://www.iso.org/sectors/it-technologies/ai

S.No	Aspect	Recommendations	Response
			serving as a benchmark for excellence. Accreditation from reputable institutions or industry-specific bodies adds credibility, providing tangible assurances of the system's robustness. By proactively pursuing these mechanisms, developers not only showcase their commitment to quality but also contribute to building trust and confidence among users and stakeholders.
6.	**Security**	New tools have to be developed that may include Input Filters and Digital Forensics Tools.	The recommendation to develop new tools such as Input Filters and Digital Forensics Tools for enhancing security in generative AI systems is promising. Implementing input filters can help in mitigating potential security risks by screening and validating input data fed into generative AI systems. Digital forensics tools can play a crucial role in post hoc analysis and investigation of security incidents involving generative AI systems. However, while these tools offer promising avenues for enhancing security in generative AI systems, it's important to uphold principles of privacy, transparency, and accountability throughout the development and implementation process, while also complying with relevant data protection laws.
7.	**Content Provenance**	• Educate Users: Help users understand the origin (provenance) of content throughout its creation process. • Verification Tools: Empower users with tools to verify the authenticity of content.	While the recommendation to focus on content provenance for generative AI systems offers valuable strategies for enhancing transparency and trustworthiness, it's important to consider some potential critiques and challenges. The recommendation

S.No	Aspect	Recommendations	Response
		- Collaboration: Work with key parties in the content lifecycle, to support the embedding and display of digital watermarks and provenance details.	to deploy digital watermarking tools to distinguish AI generated content is appreciated. However, as acknowledged in the proposal, such tools can be easily manipulated and therefore, there is a further need for development of innovative solutions. It is also recommended that cryptographic provenance solutions to track the digital content origin are deployed. However, this may result in concerns around privacy. For example, maintaining cryptographic hashes of all material created via a platform could violate the data minimization principle by unnecessarily storing excessive user data. Balancing the need for provenance with privacy considerations is crucial for ethical and effective digital content management. Furthermore, the effectiveness of content provenance measures may be limited by the proliferation of deepfake technologies and adversarial manipulation tactics, which can undermine the integrity of provenance data and erode trust in AI-generated content. Addressing this will require interdisciplinary collaboration, innovative technological solutions, and robust ethical frameworks to ensure that content provenance efforts effectively enhance transparency and accountability in the realm of generative AI systems. A collaborative push from tech companies, academic researchers, and government bodies to forge a unified set of detection benchmarks while

S.No	Aspect	Recommendations	Response
			balancing the need for independent strategies to distinguish between harmful deepfakes and legitimate synthetic media uses will be crucial. The Coalition for Content Provenance and Authenticity (C2PA)[9], a collaborative initiative led by technology leaders such as Adobe and Microsoft, is at the forefront of establishing standards for certifying media content. This consortium focuses on initiatives like the Content Authenticity Initiative[10] and Project Origin[11], which prioritize cryptographic hashing and digital watermarking to maintain content integrity. Specifically, Project Origin targets the preservation of news and information content integrity, by integrating with tools like Photoshop to ensure secure metadata preservation during editing processes. In a parallel development, DeepMind’s SynthID innovatively embeds imperceptible digital watermarks into AI-generated images or audio.[12] This technology not only enables the identification of content produced by Google's AI models but also ensures that its detectable watermark remains unaltered through subsequent image modifications. The technological strides made in content verification and integrity must be thoughtfully balanced with comprehensive policy frameworks. It's imperative that these technological advancements do not

[9] Coalition for Content Provenance and Authenticity. (2023) . Overview. C2PA. https://c2pa.org
[10] Content Authenticity Initiative. (2023). Authentic storytelling through digital content provenance. Content Authenticity. https://contentauthenticity.org
[11] Project Origin. (2023). What Origin Does. Origin Project. https://www.originproject.info
[12] SynthID. Identifying AI-generated content with SynthID. (2023, November 16). Google DeepMind. https://deepmind.google/technologies/synthid/

S.No	Aspect	Recommendations	Response
			inadvertently compromise user privacy or encroach upon the principles of free speech. To this end, The Dialogue's white paper on addressing the challenges posed by deepfakes serves as a crucial resource.[13] It provides an in-depth exploration of not only the existing and emerging technical solutions but also articulates the requisite policy measures to effectively navigate this complex landscape. This holistic approach underscores the necessity of a symbiotic relationship between technological innovation and policy development to safeguard digital content authenticity while upholding fundamental rights. On the point of educating users, while this is essential, many users may lack the technical expertise to fully understand the intricacies of how generative AI systems operate and how provenance information can be interpreted. Generative AI systems operate through complex algorithms and processes that may be challenging for the average user to comprehend without specialized knowledge or training. Therefore, efforts to educate users must be tailored to cater to diverse levels of technical literacy, employing clear and accessible language and providing practical examples to illustrate key concepts. Additionally, initiatives aimed at enhancing user understanding should incorporate interactive elements and user-friendly interfaces to facilitate engagement and comprehension.

[13] Shreya. S, and Tiwari, PB. (2024) Prevention, Detection, Reporting, and Compliance: A Comprehensive Approach towards Tackling Deepfakes in India. The Dialogue.

S.No	Aspect	Recommendations	Response
8.	**Safety and Alignment R&D**	• R&D in model safety and alignment needs to be accelerated. Global cooperation required to optimize limited talent and resources for maximum impact.	The focus on accelerating R&D in model safety highlights the framework's commitment to mitigating potential risks associated with AI development. While accelerating R&D is important, identifying the most pressing safety concerns (e.g., explainability, adversarial attacks) would help guide research efforts. R&D advancements need to be accompanied by policy and regulatory frameworks that encourage responsible AI development and deployment. Including industry, academia, and civil society in R&D discussions can ensure diverse perspectives and address real-world concerns.
9.	**AI for Public Good**	• Democratising access to technology through collaboration between industry, governments, and educational institutions. • Enhancing public service delivery through facilitating data sharing across different government agencies, access to high performance compute and other related policies. • Redesign jobs and provide upskilling opportunities and develop sustainable technologies.	The draft Framework's vision to build access to technology through collaboration is welcome. It will be crucial, however, to work on all aspects of "democratized access" including but not limited to affordability, infrastructure, digital literacy, etc. Further, strategies to ensure under-represented groups like women, minorities, rural populations, etc have equal access to opportunities needs to be outlined. On the aspect of data sharing across government agencies, it is important to establish comprehensive data governance frameworks and implement effective safety measures. These measures will be essential to ensure the protection of data privacy and maintain the integrity and security of sensitive information. Effective data governance measures like defining policies, procedures, and responsibilities for managing and utilizing data

S.No	Aspect	Recommendations	Response
			assets responsibly can be adopted. This will involve establishing clear guidelines for data collection, storage, sharing, and access, as well as outlining protocols for data usage and handling. Having said this, it's essential that data sharing requirements are not mandated; should be voluntary and is shared only on a case-to-case basis, emphasizing the importance of discretion and careful consideration in data sharing practices. Additionally, robust safety measures should be considered while sharing data like requisite security protocols, encryption techniques, access controls, and authentication mechanisms to safeguard data against unauthorized access, breaches, and misuse. Lastly, regarding the need to redesign jobs and offer upskilling opportunities, while this is undoubtedly crucial, it's equally important to ensure that upskilling programs are both affordable and accessible across all demographics. This approach is necessary to prevent the widening of skill gaps and ensure inclusivity in the workforce.

thedialogue.co

@_DialogueIndia

@TheDialogue_Official

@The-Dialogue-India

@TheDialogueIndia