ARTICLE

MIT Science Policy Review

Data and AI governance: Promoting equity, ethics, and fairness in large language models

Alok Abhishek[1,*], Lisa Erickson[2,*], and Tushar Bandopadhyay[3,*]

Edited by Swapnil Kumar and Emma Courtney

HIGHLIGHTS

- Adoption of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) is growing rapidly. Bloomberg estimates that Generative AI will become a $1.3 trillion market by 2032 [1].
- The European Union (E.U.)'s Data Governance Act (2020) and Ethics Guidelines for Trustworthy AI (2019) provide initial regulatory frameworks. Significant practical implementation challenges still remain as current data and AI governance models do not sufficiently address GenAI specific complexities, highlighting the need for frameworks that effectively balance ethical compliance with real world applicability.
- LLMs exhibit biases across many dimensions, such as gender, race and ethnicity, socioeconomic status, culture, religion, sexual orientation, disability, age, geography, political ideology, and stereotypes. Thematic analysis of biases in LLM responses show stereotypes, global and cultural dynamics, socioeconomic, and demographic related bias are prevalent in high severity and high risk responses [2].
- A data and AI governance framework that would facilitate the development of strategies to improve fairness, equity, and ethical alignment within the operational deployment of LLMs is urgently needed.

In this paper, we cover approaches to systematically govern, assess and quantify bias across the complete life cycle of machine learning models, from initial development and validation to ongoing production monitoring and guardrail implementation. Building upon our foundational work on the Bias Evaluation and Assessment Test Suite (BEATS) for Large Language Models [2], the authors share prevalent bias and fairness related gaps in Large Language Models (LLMs) and discuss data and AI governance framework to address Bias, Ethics, Fairness, and Factuality within LLMs. The data and AI governance approach discussed in this paper is suitable for practical, real-world applications, enabling rigorous benchmarking of LLMs prior to production deployment, facilitating continuous real-time evaluation, and proactively governing LLM generated responses. By implementing the data and AI governance across the life cycle of AI development, organizations can significantly enhance the safety and responsibility of their GenAI systems, effectively mitigating risks of discrimination and protecting against potential reputational or brand-related harm. Ultimately, through this article, we aim to contribute to advancement of the creation and deployment of socially responsible and ethically aligned generative artificial intelligence powered applications.

Artificial Intelligence (AI) is fundamentally reshaping the technological landscape, driven by rapid innovation, accelerated adoption, and significant financial commitments from industries worldwide. Global investment in AI technologies is projected to grow, reaching approximately $632 billion by 2028 [3, 4]. Within this broader AI expansion, investment in Generative AI (GenAI) is predicted to reach $26 billion by 2027 [5]. In the long term, the GenAI market valuation is projected to reach an estimated $1.3 trillion by 2032 [1, 6], growing at an average annual growth rate of 36% [7], further emphasizing the profound long term impact and potential. These projections highlight the scale of financial investments and underscore the accelerated pace at which generative technologies are being integrated across many sectors, signaling a paradigm shift in our use and interaction with AI in our everyday lives. As AI technologies evolve and their applications widen, the importance of comprehensive data and AI governance becomes increasingly crucial to ensure responsible and ethical development and use of AI [8, 9].

The successful ethical implementation of AI and GenAI applications, particularly in fields such as natural language processing, computer vision, and decision support systems, relies on the quality and governance of training data. The emergence of large language models (LLMs), which utilize expansive and diverse training data corpora but are prone to reflecting existing societal prejudices present in the

*Independent Researcher

[1]Email: alok@alokabhishek.ai

[2]Email: lisa.erickson@alum.mit.edu

[3]Email: tushar@kronml.com

The authors declare no conflict of interest.

training data, emphasizes the need for comprehensive data and AI governance frameworks [10]. Effective data and AI governance is essential to address the considerable risks associated with these models.

Effective regulatory frameworks are foundational to responsible development of generative AI and can help address legal, ethical, and societal impacts while facilitating innovation. Critical elements of these regulatory frameworks include comprehensive legal infrastructures, international collaboration, and adaptive governance [11]. Comprehensive legal frameworks should explicitly address intellectual property rights, accountability, transparency, and safety, ensuring alignment with human-centered principles [12, 13].

The current AI regulatory landscape focuses on AI implementation, data privacy, fairness, and transparency. Although these regulations are still evolving, they are already driving organizations across sectors, including financial services [14, 15], healthcare providers [16], legal services [17, 18], and government institutions [9, 19, 20], to navigate emerging compliance requirements. These requirements affect machine learning operations and broader data governance practices. Current AI regulations primarily focus on ensuring responsible AI development and deployment in human-impact applications. These regulations address (1) data privacy and security, (2) fairness and non-discrimination, (3) explainability, (4) transparency, and (5) environmental sustainability.

(1) Existing data privacy and data sovereignty regulations across jurisdictions impose constraints on AI deployment, particularly in relation to data collection, processing, transfer, localization, and retention. In the European Union, the General Data Protection Regulation (GDPR) (1) mandates stringent consent and purpose limitation requirements. In the United States, the California Consumer Privacy Act (CCPA) (2) enforces consumer rights to access, delete, and opt out of data sharing. Beyond these, many countries across the globe have enacted national frameworks to assert data sovereignty and protect citizen privacy.

For instance, China has implemented a comprehensive set of laws including the Cybersecurity Law (6), the Data Security Law (7), and the Personal Information Protection Law (PIPL) (8), which impose strict data localization and cross-border data transfer restrictions. Similarly, Russia's Federal Law on Personal Data (9) requires that personal data of Russian citizens be stored and processed within national borders.

India's Digital Personal Data Protection Act (DPDPA) (5) and Reserve Bank of India's data localization mandates (4) reflect growing emphasis on data sovereignty in South Asia. In Latin America, Brazil's Lei Geral de Proteção de Dados (LGPD) (10) parallels GDPR in its scope and enforcement mechanisms. Australia's Privacy Act (11) continues to evolve to address digital and AI era privacy concerns.

In the Middle East, the United Arab Emirates has adopted multiple frameworks, including the Federal Decree Law No. 45 of 2021 (12), the Dubai International Financial Centre (DIFC) Data Protection Law No. 5 of 2020 (13), and the Abu Dhabi Global Market (ADGM) Data Protection Regulations (14). Other Gulf states such as Saudi Arabia (15), Qatar (16,17), Bahrain (18), Oman (19), and Kuwait (20) have introduced similar data protection laws that emphasize both individual rights and national data governance.

Collectively, these regulations are increasingly influencing how firms design, develop, and deploy AI and data platforms. Compliance with these frameworks now necessitates a deeper alignment of AI lifecycle governance with jurisdiction specific data handling requirements, thus adding complexity to cross-border AI scalability and innovation.

(2) Fairness and non-discrimination regulations, including the Equal Credit Opportunity Act (3) and the more targeted proposed E.U. AI Act (22,23), have strict requirements that AI systems remain free of biases related to race, gender, or religion. The E.U. AI act seeks to impose strict rules for high-risk AI applications to make sure customers are treated fairly. Fairness requirements creates difficulties in adoption of Gen AI as, GenAI models are trained on common internet data which might reinforce societal biases, leading to unfair outcomes [2, 10, 21].

(3) Explainability and transparency requirements, like those outlined by the Basel Committee on Banking Supervision [22] and Model Risk Management (27), demand that AI models be interpretable, auditable, and fair. These create difficulties for generative AI models, which often function as a "black box," and make it difficult to explain outputs or meet interpretability requirements.

(4) The European Union's Data Governance Act (2022) (2) and the Ethics Guidelines for Trustworthy AI (2019) (23) have laid down preliminary regulatory frameworks. These regulations focus on increasing trust in data sharing, and development of lawful, ethical, and robust AI applications.

(5) While regulations such as the European Energy Efficiency Directive (EED) (24), which mandates data centers and technology intensive enterprises to optimize and transparently report their energy consumption, and standards like International Organization (ISO) 14001 (Environmental Management) [23], which promotes a systematic approach to managing environmental impacts, directly shape technology infrastructure operations and, by extension, AI operations, broader governance frameworks and policy guidelines such as Organization for Economic Co-operation and Development (OECD) AI Principles (25) and the E.U. AI Act (22) explicitly integrate sustainability into the frameworks. These AI specific guidelines encourage resource efficient training and inference processes, minimal environmental footprints, and transparent reporting of AI models' carbon

emissions, thereby aligning AI development with broader environmental sustainability objectives.

Geopolitical considerations and AI sovereignty are also emerging areas of regulatory focus, with potential impacts for AI regulation. Emerging regulations may also address geopolitical concerns. National security and international competition in AI technology are driving trade restrictions, data localization, and sovereignty laws that could reshape global AI development (26).

While these regulatory frameworks provide a foundation for the security and privacy of data and ethical and responsible AI development, translating these high-level governance guidelines into effective real-world operational governance practice remains challenging. A critical gap lies in addressing the complex and nuanced ethical dilemmas and biases prominent in LLMs, which are trained on extensive but generic internet datasets prone to perpetuating societal biases and inequities. Despite regulatory efforts toward fairness, transparency, and explainability, practical implementation often struggles to keep pace with rapidly evolving AI capabilities. The scale and complexity of contemporary LLMs magnify these challenges, requiring adaptive and proactive methods to detect, quantify, and mitigate bias and ethical shortcomings. Therefore, a closer examination of specific bias-related shortcomings in leading LLMs is essential to identify areas where governance frameworks must evolve, bridging the gap between regulation and real-world ethical AI performance.

Ethics and bias related shortcomings in leading large language models

A key focus in the ongoing efforts to improve AI governance is understanding and mitigating bias in GenAI systems. Bolukbasi et al., in their 2016 paper *"Man is to computer programmer as woman is to homemaker? debiasing word embeddings"*, demonstrated that word embedding models trained on common internet data like Google News encode and even amplify gender stereotypes to a dangerous extent [10]. This work showed how statistical correlations in training data can reinforce harmful societal prejudices. This paper laid the foundation for bias detection and drove impetus to reduce bias in early stage Natural Language Processing (NLP) systems.

Parrish et al., in the paper *"BBQ: A Hand-Built Bias Benchmark for Question Answering"*, introduce the Bias Benchmark for Question Answering (BBQ), a dataset designed to evaluate how social biases manifest in the outputs of question-answering (QA) models [21]. This study highlighted that Natural Language Processing (NLP) models often reproduce harmful stereotypes, leading to biased outputs.

In our recently published ArXiv paper *"BEATS: Bias Evaluation and Assessment Test Suite for Large Language Models"*, we introduced the Bias Evaluation and Assessment Test Suite (BEATS), a novel framework for evaluating Bias, Ethics, Fairness, and Factuality in Large Language Models (LLMs) [2]. Utilizing this framework, we demonstrated that 37.65% of outputs generated by industry leading large language models contained some form of bias. Furthermore, 33.7% of responses have either a high or medium level of bias severity or potential impact, highlighting a substantial risk of using these models in critical decision making systems.

Biases in LLMs span numerous dimensions, including stereotypes, cultural dynamics, individual and community biases. Biases related to age and gender form a core theme of reoccurring biases in the output of LLMs. Disability accommodations, gender roles, religious beliefs, economic disparities, and cultural influences are also prevalent. Prominent themes included systemic barriers, intergenerational technological adaptation, and socioeconomic challenges, highlighting complex interactions among demographic, cultural, and economic factors influencing LLM bias.

Studies such as Slattery et al. [24] and Weidinger et al. [25] underscore that the deployment of AI across high-stakes domains such as healthcare, law, finance, and public infrastructure introduces substantial risks that are frequently underestimated and inadequately governed. These risks include misinformation, systemic bias, privacy violations, weaponization, and regulatory failures and can arise from both human actions and autonomous AI behavior. Large language models have shown the potential to cause significant ethical and social harm, including discrimination, toxicity, information leakage, and automation-driven socioeconomic disruptions [2, 25]. These harms have been observed in research studies as well as in the real world and tend to disproportionately affect marginalized communities due to biased training data and uneven model performance.

These findings clearly demonstrate that large language models exhibit biases across many dimensions, such as gender, race and ethnicity, socioeconomic status, culture, religion, sexual orientation, disability, age, geography, political ideology, and stereotypes. Therefore, incorporating LLMs into critical decision making applications, especially in domains such as healthcare [8, 26], legal services [17, 18], finance [27, 28], and governance [9, 19], raises major risks and ethical concerns due to these inherent biases, potentially exacerbating systemic inequities. Given the high stakes repercussions of such biases, it is imperative to establish a rigorous governance framework that leverages statistical methodologies to systematically assess, mitigate, and manage biases. Such a governance framework would facilitate the development of strategies to improve fairness, equity, and ethical alignment within the operational deployment of LLMs.

Overview of data and AI lifecycle

Data governance in AI involves the systematic management of data assets throughout their lifecycle from acquisition and storage to processing, deployment, and decommissioning. The lifecycle of AI systems is a structured and iterative process rather than a linear pipeline, aligning

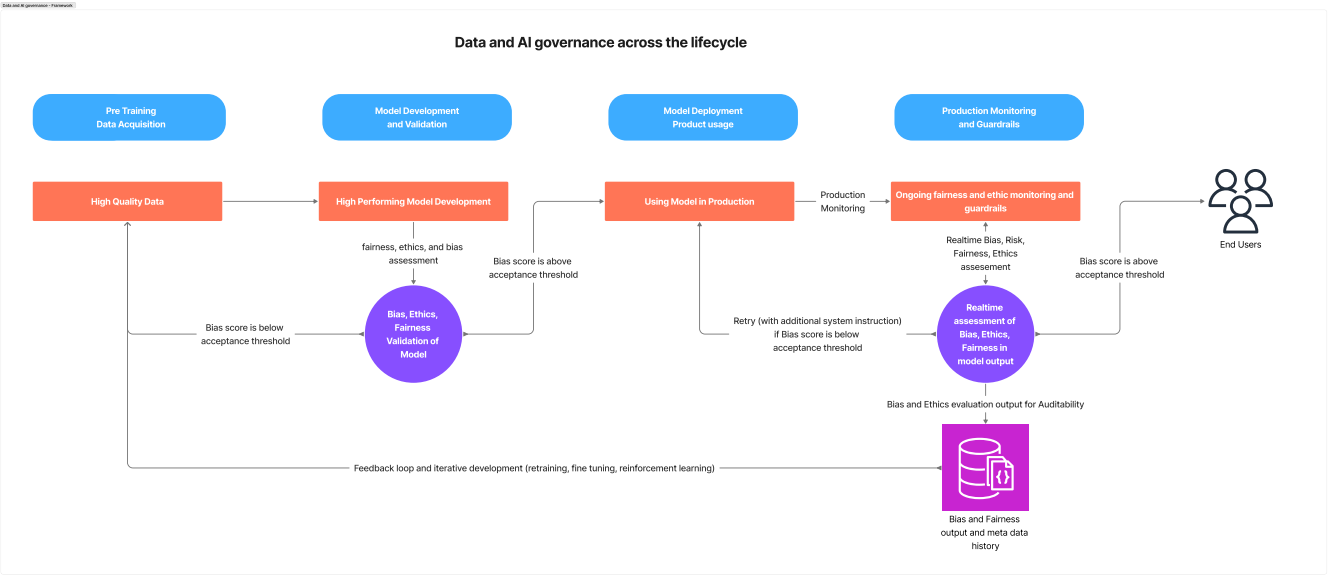

Figure 1: System design of data and AI governance across the AI life cycle. The bias evaluation is performed as part of overall model evaluation before deploying the model in products and as an ongoing guardrail during model inference responses in production.

with contemporary frameworks such as the CDAC (Collect, Design, Assess, and Consume) model proposed by Silva et al. and calls to revise traditional AI lifecycle models to accommodate evolving operational demands and ethical considerations by Haakman et al. [29, 30].

The machine learning lifecycle spans five interdependent phases. It begins with data collection and preparation, where high-quality datasets are acquired from diverse sources and undergo rigorous curation, validation, auditing, and preprocessing to uphold ethical standards, mitigate bias, and support feature engineering. In the model development phase, appropriate algorithm selection, baseline establishment, hyperparameter optimization, fairness assessments, and explainability techniques are employed to develop performant and trustworthy models.

Deployment introduces the need for secure, scalable infrastructure and often involves containerizing models, exposing them through inference APIs, and implementing governance controls and guardrails that ensure compliance with organizational and regulatory standards. Post-deployment, continuous monitoring and maintenance are essential to detect model drift, track performance, and ensure ongoing fairness and ethical alignment for embedded governance checkpoints.

Finally, iterative improvement and retirement processes are driven by user feedback and production data insights. Models are refined continuously or retired systematically, with secure archival of model artifacts to ensure auditability and reproducibility. This holistic perspective, supported by frameworks like CDAC, highlights the iterative, sociotechnical nature of AI system lifecycles and the importance of aligning them with evolving best practices in AI ethics, transparency, and accountability.

Governance across the data and AI lifecycle

A comprehensive data and AI governance approach to mitigating biases and ethical harms in generative AI models must systematically integrate governance practices across the entire data and machine learning (ML) lifecycle.

It begins with the data collection and acquisition phase, where strategies include emphasizing source verification to ensure data diversity, reliability, and compliance with privacy regulations such as GDPR and CCPA. Additionally, regular demographic diversity and representativeness audits can be conducted, complemented by rigorous consent management and anonymization protocols for safeguarding sensitive information.

During the data preprocessing and labeling stage, bias detection techniques can be deployed to identify and correct systemic imbalances using statistical methods and normalization strategies. Transparent and standardized labeling protocols are implemented by diverse teams to minimize subjective and cultural biases. Comprehensive documentation, including meticulous version control and detailed data sheets, ensures data lineage traceability and accountability.

To ensure that AI systems are fair, transparent, and responsible, model development and training governance incorporate several critical practices. These include selecting fairness-aware algorithms that help minimize bias, conducting tests against misleading or adversarial inputs, and embedding ethical considerations throughout the design and development lifecycle. Independent ethics and bias review boards, along with formal fairness impact assessments, play a key role in maintaining accountability across the development process. Explainable AI (XAI) techniques, which make AI output easier to understand, are utilized to make model predictions more interpretable [31]. Leading XAI methods, such as Shapley

Additive Explanations (SHAP) [32], Local Interpretable Model-Agnostic Explanations (LIME) [33], Partial Dependence Plots (PDPs) [34], and Counterfactual Explanations [35], help clarify how models arrive at specific results. These tools enhance transparency by documenting model assumptions, identifying potential limitations, and providing insights into the rationale behind predictions [26].

Validation and testing governance emphasize continual evaluation against established fairness metrics, employing cross group performance analysis to detect potential biases. Rigorous robustness and factual consistency tests, including counterfactual and out of sample evaluations, are performed alongside human in the loop verification processes to address edge cases and incorporate expert insights, fostering a feedback driven approach to bias mitigation. Human oversight remains crucial for accountability, emphasizing human in the loop models to ensure responsible AI operation.

In the model deployment and monitoring phase, the framework maintains continuous fairness observability through real-time dashboards and routine audits. Ethical feedback mechanisms are strategically designed to prevent the reinforcement of biases during model updates. Incident response protocols provide structured mechanisms for addressing ethical breaches, ensuring timely resolution and user redress.

Finally, lifecycle governance emphasizes policies and practices that align with global AI regulations and ethical standards, supported by transparent stakeholder engagement, such as publicly accessible models and data cards. Continuous improvement initiatives integrate learning from ethical incidents, promoting an organizational culture deeply committed to ethical AI practices through regular training and external audits.

Application of the governance approach across the data and AI life cycle

Figure 1 shows a governance framework approach that covers the whole AI life cycle. This data and AI governance framework designed for real-world applications, enabling real-time assessment of LLM responses. This capability is essential for ensuring safe, responsible AI deployments that proactively mitigate discrimination risks and prevent potential brand or reputational damage.

This governance approach for data and AI is designed to integrates into two core phases of data and AI governance: model development and production guardrails. During the model development and validation phase, candidate models undergo evaluation for bias. Models achieving scores above the predefined acceptance threshold are approved for deployment into production environments. Conversely, models falling below this threshold trigger a retraining process, prompting a review of data curation practices to ensure training datasets do not reinforce societal inequities and are adequately representative of diverse global cultures.

In the production phase, this framework facilitates real-time assessments of model outputs, proactively identifying and mitigating ethical and fairness risks. Outputs exceeding acceptance thresholds proceed smoothly to subsequent stages, such as reaching the end users. However, if outputs falling below the acceptance thresholds the guardrails prevents the high risk responses from advancing further. The framework initiates a retry mechanism supplemented by additional prompt instructions, regenerating responses that align with ethical standards and fairness criteria.

This governance approach also incorporates an adaptive feedback mechanism, enabling continuous iterative improvements through retraining, fine-tuning, and reinforcement learning. This iterative cycle ensures that models learn from real world data insights and user feedback, facilitating continuous refinement and enhanced ethical alignment.

Limitations

While the previously discussed data and AI governance framework provides comprehensive guidance for mitigating bias and ethical challenges across the lifecycle of LLM-powered applications, several limitations should be acknowledged:

1) *Dynamic regulatory and ethical landscape*: The governance framework described previously aligns with current regulatory guidelines and ethical considerations. However, the rapidly evolving global AI regulatory and fast evolving frontier LLM landscape implies that ongoing adjustments and refinements will be necessary. Governance approaches must remain flexible and adaptive to evolving regulatory, societal, and technological developments.

2) *Scope and generalizability of the framework*: The framework discussed in this paper is primarily designed for governance in generative AI and LLM contexts. Although many governance practices can generalize to other AI domains, the specific methodologies and approaches may require adaptation for different types of AI systems, particularly those employing structured or multimodal data or those involving reinforcement learning or realtime interactive AI.

3) *Limitations of bias measurement methods*: While the BEATS framework [2] provides structured evaluations of bias, fairness, ethics, and factuality, it uses LLMs as a judge paradigm where judge and generative models share similar training data, which is predominantly english and western culture centric. This could lead to a self-reinforcing mechanism, where a lack of global and diverse training data sets leads to a lack of sensitivity towards underrepresented or nondominant global viewpoints [36, 37]. Therefore, there is a risk of evaluation scores representing fairness and ethical alignment, which are not global in nature.

Conclusion

Application of Generative AI, significantly accelerated

by the introduction of transformer architecture by Vaswani et al. [38], has fundamentally reshaped the landscape of applied artificial intelligence, influencing numerous industries ranging from finance and healthcare to governance and everyday applications worldwide. As these generative AI powered applications become deeply embedded in critical decision making processes, their societal implications have grown, raising the need for heightened scrutiny of ethical and regulatory compliance.

A growing body of scholarly research, combined with evolving regulatory frameworks, including the Equal Credit Opportunity Act (3), transparency and explainability mandates from the Basel Committee on Banking Supervision [22], Model Risk Management guidelines (27), and fairness provisions proposed under the E.U. AI Act (22,23) has significantly contributed to development of fairer, more transparent, and accountable AI practices. Despite these regulatory advancements, research by Bolukbasi et al. [10], Parrish et al. [21], and Abhishek et al. [2] clearly highlights persistent risks that Large Language Models (LLMs) perpetuate existing societal biases, potentially exacerbating systemic inequalities.

To bridge this gap between regulatory intent and practical application, we discuss a comprehensive data and AI governance framework. This governance approach systematically measures, monitors, and mitigates biases and ethical concerns, integrating fairness assessments, transparency, and privacy measures across the entire AI lifecycle. By proactively aligning operational practices with compliance requirements, including GDPR (1), the E.U. Data Governance Act (2), and sustainability oriented guidelines such as ISO 14001 [23] and OECD AI Principles (25), organizations can enhance their AI systems' ethical alignment and regulatory adherence.

The adaptive nature of the discussed governance framework equips organizations to effectively navigate the dynamic regulatory and geopolitical landscapes, including data sovereignty and international AI governance concerns. Continuous monitoring and iterative enhancements embedded within this framework not only facilitate compliance but also foster stakeholder trust, mitigate reputational risks, and ensure the sustainable, equitable, and ethical deployment of AI technologies. This structured governance model advances the overarching goal of responsibly adopting the transformative potential of generative AI to create transparent, fair, and ethically aligned GenAI powered applications that genuinely benefit society.

Future research directions

With the larger goal of contributing to the development of fairer LLMs that do not perpetuate societal biases and are suitable for use in critical decision making systems, we intend to continue future research in this area. To further extend the impact of the discussed governance framework, we outline several opportunities for future research:

1) *Empirical validation and framework optimization:* Empirical validation of the effectiveness of this governance framework across various industries would significantly strengthen the credibility and practical applicability of the approach. Detailed case studies and quantitative analyses will help refine the framework, identifying opportunities for optimization and tailored adaptations across diverse operational environments.
2) *Exploration of multimodal GenAI governance:* Given the increasing prevalence of multimodal AI systems, future research to extend and refine the governance strategies that encompass multimodal data (e.g., images, video, and audio combined with text) will contribute greatly to the development of strategies for ethical and equitable GenAI applications.
3) *Development of interactive governance tools:* We intend to develop practical, user-friendly tools and interfaces based on the Bias Evaluation and Assessment Test Suite (BEATS) [2]. These tools will facilitate real-time assessment, bias monitoring, and interpretability of model outputs, enabling practitioners and researchers to more readily identify, quantify, and mitigate biases in operational LLM powered GenAI applications.

Acknowledgments

The authors extend sincere gratitude to the *MIT Science Policy Review (SPR)* for the opportunity to contribute to this important discourse. We are especially thankful to *Swapnil Kumar*, *Emma Courtney*, and *Audrey Bertin* for their support and insightful guidance throughout the editorial process. We also deeply appreciate the thoughtful and constructive feedback provided by the peer reviewers.

Citation

Abhishek, A., Erickson, L. & Bandopadhyay, T. Data and AI governance: Promoting equity, ethics, and fairness in large language models. *MIT Science Policy Review* **6**, 139-146 (2025). https://doi.org/10.38105/spr.1sn574k4lp.

Legislation cited

(1) European Union. General Data Protection Regulation (GDPR), Regulation (E.U.) 2016/679 (2016). Online: https://eur-lex.europa.E.U./eli/reg/2016/679/oj
(2) California State Legislature. California Consumer Privacy Act (CCPA) (2018). Online: https://oag.ca.gov/privacy/ccpa
(3) United States Congress. Equal Credit Opportunity Act (ECOA) (1974). Online: https://www.consumerfinance.gov/rules-policy/regulations/1002/
(4) Reserve Bank of India. Storage of Payment System Data (2018). Online: https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=11244
(5) Parliament of India. Digital Personal Data Protection Act (2023). Online: https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf
(6) National People's Congress of China. Cybersecurity Law (2016). Online: http://www.cac.gov.cn/2016-11/07/c_1119867116.htm
(7) National People's Congress of China. Data Security Law (2021). Online: http://www.npc.gov.cn/npc/c30834/202007/3b5d6944cbb74d3b9b8e5a4f6f3097e4.shtml
(8) National People's Congress of China. Personal Information Protection Law (PIPL) (2021). Online: http://www.npc.gov.cn/englishnpc/c23934/202111/2b1f0e1a2e3e4a2b9b8e5a4f6f3097e4.shtml
(9) Russian Federation. Federal Law on Personal Data (2006). Online: https://en.wikipedia.org/wiki/Data_protection_(privacy)_laws_in_Russia
(10) Brazil. Lei Geral de Proteção de Dados Pessoais (LGPD) (2018). Online: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
(11) Parliament of Australia. Privacy Act 1988. Online: https://www.legislation.gov.au/Details/C2023C00210
(12) United Arab Emirates. Federal Decree-Law No. 45 of 2021 on the Protection of Personal Data. Online: https://u.ae/en/about-the-uae/digital-uae/data/data-protection-laws
(13) Dubai International Financial Centre. Data Protection Law No. 5 of 2020. Online: https://www.difc.ae/business/laws-regulations/data-protection/
(14) Abu Dhabi Global Market. Data Protection Regulations 2021. Online: https://www.adgm.com/operating-in-adgm/doing-business/data-protection
(15) Saudi Data and AI Authority. Personal Data Protection Law (PDPL) (2021). Online: https://sdaia.gov.sa/en/Research/Pages/DataProtection.aspx
(16) State of Qatar. Law No. 13 of 2016 on Personal Data Protection. Online: https://www.motc.gov.qa/en/documents/document/law-no-13-2016-concerning-personal-data-privacy-protection
(17) Qatar Financial Centre. Data Protection Regulations 2021. Online: https://www.qfc.qa/en/operating-in-the-qfc/legal-and-regulatory/data-protection
(18) Kingdom of Bahrain. Law No. 30 of 2018 Promulgating the Personal Data Protection Law. Online: https://www.legalaffairs.gov.bh/Media/LegalPDF/K3001.pdf
(19) Sultanate of Oman. Royal Decree No. 6/2022 Promulgating the Personal Data Protection Law. Online: https://www.mola.gov.om/Download.aspx?Path=royal/2022-0006.pdf
(20) State of Kuwait. Data Privacy Protection Regulation (2023). Online: https://www.citra.gov.kw/sites/en/Pages/Data-Privacy-Protection.aspx
(21) European Union. Data Governance Act (2022). Online: https://digital-strategy.ec.europa.E.U./en/policies/data-governance-act
(22) European Union. Artificial Intelligence Act (2024). Online: https://artificialintelligenceact.E.U./
(23) European Union. Ethics Guidelines for Trustworthy AI (2019). Online: https://digital-strategy.ec.europa.E.U./en/library/ethics-guidelines-trustworthy-ai
(24) European Union. Energy Efficiency Directive (2015). Online: https://energy.ec.europa.E.U./topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en
(25) Organization for Economic Co-operation and Development (OECD). Recommendation of the Council on Artificial Intelligence (2019). Online: https://oecd.ai/en/ai-principles
(26) UK Government. International AI Safety Report (2025). Online: https://www.gov.uk/government/publications/international-ai-safety-report-2025
(27) United States. Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management (2011). Online: https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html

References

[1] Bloomberg. Generative AI to become a $1.3 trillion market by 2032, research finds (2023). Online: https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/.
[2] Abhishek, A., Erickson, L. & Bandopadhyay, T. BEATS: Bias evaluation and assessment test suite for large language models. *arXiv* (2025). https://arxiv.org/abs/2503.24310.
[3] International Data Corporation (IDC). Worldwide spending on artificial intelligence forecast to reach $632 billion in 2028 (2023). Online: https://www.idc.com/getdoc.jsp?containerId=prUS52530724.
[4] S&P Global. Generative AI market forecasts revised upward to $52.2b by 2028 (2023). Online: https://www.spglobal.com/marketintelligence/en/news-insights/research/generative-ai-market-forecasts-revised-upward-to-52-2b-by-2028.
[5] International Data Corporation (IDC). Generative AI spending to reach $26 billion by 2027 (2023). Online: https://www.idc.com/getdoc.jsp?containerId=prAP52048824.
[6] Precedence Research. Generative AI market size, share, and trends 2024 to 2033 (2023). Online: https://www.precedenceresearch.com/generative-ai-market.
[7] Forrester. Spend on generative AI will grow 36% annually to 2030 (2023). Online: https://www.forrester.com/blogs/spend-on-generative-ai-will-grow-36-annually-to-2030/.
[8] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns (N Y)* **2**, 100347 (2021). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8515002/.
[9] Alhosani, K. & Alhashmi, S. M. Opportunities, challenges, and benefits of ai innovation in government services: a review. *Discover Artificial Intelligence* **4** (2024). https://link.springer.com/article/10.1007/s44163-024-00111-w.
[10] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv* (2016). https://arxiv.org/abs/1607.06520.
[11] Luna, J., Tan, I., Xie, X. & Jiang, L. Navigating governance paradigms: A cross-regional comparative study of generative AI governance processes & principles. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 917–931 (2024). http://dx.doi.org/10.1609/aies.v7i1.31692.
[12] de Almeida, P. G. R., dos Santos, C. D. & Farias, J. S. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology* **23** (2021). https:

//link.springer.com/article/10.1007/s10676-021-09593-z.

[13] Li, J., Cai, X. & Cheng, L. Legal regulation of generative AI: a multidimensional construction. *International Journal of Legal Discourse* **8**, 365–388 (2023). https://www.degruyterbrill.com/document/doi/10.1515/ijld-2023-2017/html.

[14] Lee, J. Access to finance for artificial intelligence regulation in the financial services industry. *European Business Organization Law Review* **21** (2020). https://link.springer.com/article/10.1007/s40804-020-00200-0.

[15] Vuković, D. B., Dekpo-Adza, S. & Matović, S. Ai integration in financial services: a systematic review of trends and regulatory challenges. *Humanities and Social Sciences Communications* **12** (2025). https://link.springer.com/article/10.1007/s40804-020-00200-0.

[16] Palaniappan, K., Lin, E. Y. T. & Vogel, S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. In *Healthcare*, vol. 12, 562 (2024). https://www.mdpi.com/2227-9032/12/5/562.

[17] Kapoor, S., Henderson, P. & Narayanan, A. Promises and pitfalls of artificial intelligence for legal applications (2024). https://arxiv.org/abs/2402.01656, 2402.01656.

[18] Magesh, V. *et al.* Hallucination-free? assessing the reliability of leading ai legal research tools (2024). https://arxiv.org/abs/2405.20362, 2405.20362.

[19] Mellouli, S., Janssen, M. & Ojo, A. Introduction to the issue on artificial intelligence in the public sector: Risks and benefits of ai for governments. *Digit. Gov.: Res. Pract.* **5** (2024). https://dl.acm.org/doi/full/10.1145/3636550.

[20] Yigitcanlar, T., Agdas, D. & Degirmenci, K. Artificial intelligence in local governments: perceptions of city managers on prospects, constraints and choices. *AI & SOCIETY* **38** (2023). https://link.springer.com/article/10.1007/s00146-022-01450-x.

[21] Parrish, A. *et al.* BBQ: A hand-built bias benchmark for question answering. *Findings of the Association for Computational Linguistics: ACL 2022* **2022** (2022). https://aclanthology.org/2022.findings-acl.165/.

[22] Goodhart, C. *The Basel Committee on Banking Supervision: A History of the Early Years, 1974–1997* (Cambridge University Press, 2011). https://doi.org/10.1017/CBO9780511996238.

[23] International Organization for Standardization. Iso 14001 - environmental management systems — requirements with guidance for use. https://www.iso.org/standard/60857.html (2015).

[24] Slattery, P. *et al.* The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv* (2025). https://arxiv.org/abs/2408.12622.

[25] Weidinger, L. *et al.* Ethical and social risks of harm from language models (2021). https://arxiv.org/abs/2112.04359.

[26] Elshawi, R., Al-Mallah, M. H. & Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inform. Decis. Mak.* **19**, 146 (2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6664803/.

[27] Qureshi, N. I., Choudhuri, S. S., Yaramala Nagamani, R. A. V. & Shah, R. Ethical considerations of ai in financial services: Privacy, bias, and algorithmic transparency. *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* **1** (2024). https://ieeexplore.ieee.org/abstract/document/10616483.

[28] Zhang, Y. & Zhou, L. Fairness assessment for artificial intelligence in financial industry (2019). https://arxiv.org/abs/1912.07211, 1912.07211.

[29] Silva, D. D. & Alahakoon, D. An artificial intelligence life cycle: From conception to production. *Patterns* **3** (2022). https://doi.org/10.1016/j.patter.2022.100489.

[30] Haakman, M., Cruz, L., Huijgens, H. & van Deursen, A. Ai lifecycle models need to be revised. an exploratory study in fintech. *Empirical Software Engineering* **26** (2022). https://doi.org/10.1007/s10664-021-09993-1.

[31] Wikipedia. Explainable artificial intelligence (2017). Online (Accessed 12-Apr-2025): https://en.wikipedia.org/wiki/Explainable_artificial_intelligence.

[32] Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv* (2017). https://arxiv.org/abs/1705.07874.

[33] Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should i trust you?": Explaining the predictions of any classifier. *arXiv* (2016). https://arxiv.org/abs/1602.04938.

[34] Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M. & Bischl, B. Explaining hyperparameter optimization via partial dependence plots. *arXiv* (2022). https://arxiv.org/abs/2111.04820.

[35] Verma, S., Dickerson, J. & Hines, K. Counterfactual explanations for machine learning: Challenges revisited. *arXiv* (2021). https://arxiv.org/abs/2106.07756.

[36] Ye, J. *et al.* Justice or prejudice? Quantifying biases in LLM-as-a-judge. *arXiv* (2024). https://arxiv.org/abs/2410.02736.

[37] Zheng, L. *et al.* Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv* (2023). https://arxiv.org/abs/2306.05685.

[38] Vaswani, A. *et al.* Attention is all you need. *arXiv* (2023). https://arxiv.org/abs/1706.03762.