EXPLORING CHALLENGES IN THE TRANSPARENCY, BIAS AND POLICY ALIGNMENT IN GENERATIVE AI

Dr. T. Premalatha[1], Ms. A. S. Nithya[2]

[1]*Data Science Department, Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamilnadu, India.*

premalathat@skacas.ac.in,

[2]*Artificial Intelligence & Machine Learning Department, Sree Saraswathi Thyagaraja College of Arts and Science, Pollachi, Tamilnadu, India,*

nithyaas@stc.ac.in

Abstract

Generative artificial intelligence is changing the game in the industries that it touches, paving the way for new applications (text or image generation, etc). However, it also faces a number of challenges, including (but not limited to) transparency in regards to data use, bias reduction in the generated output, resistance to adversarial attacks, data privacy provisions, and ability to scale to constantly evolving policy landscapes. This paper introduces novel contributions to tackle these issues, including an adaptive debiasing framework that dynamically updates training datasets based on demographic and cultural diversity, and a dual-layer defense mechanism combining adversarial training with federated learning to enhance security without compromising user privacy. All these advances, together with an in-depth exploration of the current state of affairs with respect to the challenges and their solutions, such as fairness-aware techniques, differential privacy, and federated learning, have a cumulative effect in establishing principles from which AI systems can be built that are robust, ethical and compliant with user and regulator needs.

Keywords: Generative AI, Transparency, Bias, Privacy, Collaboration

1. INTRODUCTION

Generative AI has been developing at a speed that even could be unbelievable, changing tasks such as writing of articles and image creation to fully automated ones. Although these achievements are exciting, they also pose serious problems in the filed of cybersecurity and privacy. Problems regarding, e.g., information being transparent in terms of how it is collected and used, biases in the generated outputs, robustness to adversarial attacks, and privacy of confidential data have emerged as critical issues.

This paper gives a new perspective to these difficulties by proposing, an adaptive framework to overcome the issues of bias. Whereas static datasets have been used, the proposed method dynamically updates the training data to take account of ongoing cultural and demographic changes, resulting in fairer AI outputs. Second, we propose a strong dual-layer protection strategy based on adversarial learning integrated with federated learning, which provides better protection against security attack without losing user privacy. Since the rapid development of Generative AI, bridging the gap of these approaches with new legislation and regulations is imperative to allow responsible and ethical applications. This paper explores these questions, proposing novel solutions and actionable ideas.

2. LITERATURE SURVEY

2.1 Transparency

Transparency in Generative AI is more than a buzzword, it's about the ability, to users, developers, and regulators to understand clearly how these systems function, what data they are trained on and the potential risks that they pose. It is one of the most critical determinants for building trust, establishing accountability, and following legal frameworks like GDPR [2]. Techniques like differential privacy have developed for the purpose of privacy within the use of personal data while allowing the use of this data for training AI models. Besides, the emergence of the tools, such as Google's Model Cards and OpenAI's system reports, is establishing a path to better understanding of the capabilities, risks, and limits of AI [2, 3]. Yet, transparency isn’t just a technical challenge—it’s an ongoing responsibility. However, bias in AI systems such as has to be recognized and fought against transparently. Ethical transparency is also seen in disclosing vulnerabilities and their resolutions in another way, such as OpenAI's taking an early stance on the misuse risks [4, 5]. Also, watermarking AI-generated content is becoming more popular as a solution to cope with problems such as deepfakes [6]. Keeping realistic awareness reports current as technologies change is an enormously difficult problem, but necessary if trust in AI systems is ever to be earned [7, 8].

2.2 Bias Management

The lack of "bias" in artificial intelligence (AI) is an ongoing issue arising from the nature of training data sets, from the model they are based on, and the application environment. It is a significant issue for fair and satisfactory (robust) systems [9]. Heterogeneous and representative datasets are a base, but they are still only the first step. Recently developed algorithms have recently appeared that can both unveil and (gradually) subdue this bias by continuously (monitored and) adapting the model over the course of training [10, 11, 12].

Human oversight is also invaluable. Since this panel review of reviewers (and not of automated technologies as e.g., textmine) entails the potential for blunders that are not detected by automated reviewers (as, e.g., textmine) and also involves the potential for some accountability to be on the reviewers [16]. However, it is equally important to document existing record of the design transparency of such systems and the platforms that they are built upon, including limitations of the data sources that they use [17]. Despite these efforts, managing bias is far from straightforward. Fairness is a culturally and demographically determined construct for that reason, an evolving and dynamic issue (18).

2.3 Attack Resilience

With the progress of Generative AI platform development, Generative AI is attracting more and more attention as a target of malicious attack. Representing adversarial inputs to data leaks, these attacks point out the necessity for defenses [19]. Adversarial training and federated learning are proposed as techniques for system protection and respecting user privacy [19, 20].In practice (e.g., Google's federated learning system), it has been proven that these techniques can protect user privacy without performance loss [20]. Nevertheless, the tradeoff between security and performance is still tenuous. Attackers are continuously looking for new ways to take advantage of vulnerabilities and there is a necessity for researchers to learn and evolve [21,22].

2.4 Data Privacy Techniques

User data privacy is a pillar of ethical AI. Methods such as differential privacy, homomorphic encryption, can maintain privacy on the application level of data itself while the output of the data is core to the formation of a model [20]. One promising pathway to lower the risk of disclosure of data is federated learning in which data resides on the user device [20]. Moreover, synthetic data generation has also been proposed as a method for emulating or simulating behavior in the real world in a privacy-unsecured way [20].

Despite these advances, challenges remain. The quest for an appropriate level of privacy vs. performance is a difficult dilemma, particularly in the field of applications [36]. There is still a fascinating area where there is still room to improve, that is, the seamless integration of international borders within the control (that is, providing privacy controls) is a hot topic for research in the AI community.

2.5 Policy Alignment

Generative AI has changed the art and practice of content creation and consumption but it will nonetheless need to be woven into robust policy frameworks. Regulation (e.g., GDPR and CCPA) sets standards for transparency, consent and accountability, but reaching worldwide normative practice is a daunting task [23]. Governments and agencies are looking to provide a clearer definition of how AI systems are designed, and the data that they are grounded in, and how to encourage ethical principles in their design [24].

The challenge isn't just about compliance—it's about fostering trust. Inroads such as the EU's AI Act, as well as corporate standards from Google and Facebook illustrate the international call for ethical use of AI [26]. At the same time, the policy shall keep pace with the rapid development of technology and adequately safeguard against deviated and unfair practices.

3. INSIGHTS

3.1 ENHANCING TRANSPARENCY IN AI

Obstacles including the lack of a standardized data collection process, the unclear reasons underlying the deployment of AI, as well as the different regulatory requirements continue to be obstacles for achieving full transparency [27]. In order to mitigate these, this paper highlights the implementation of adaptive audit mechanisms with continuous feedback of data consumption and system activity thus ensuring higher accountability. Next generation of tools such as model cards and transparency reports can be still improved in order to integrate dynamic updates to reflect real-time risks and limitations, thus closing the gap between user needs and system abilities [28].

3.2 MANAGING BIAS IN AI

Bias in AI can often be traced back to static training datasets that do not represent real world situations [29]. In this work, an adaptive debiasing framework is proposed that can update the datasets in response to changing demographic and cultural diversity of the population, to increase the inclusiveness of artificial intelligence systems. Furthermore, fairness-aware algorithms are emphasized as important means by which balanced outputs can be handled during both the model training and post-her treatment [30, 31]. Repeat auditing and multidisciplinary review groups are still crucial to help expose and avoid biases [32].

3.3 STRENGTHENING ROBUSTNESS AGAINST ATTACKS

Generative AI are threatened by an increasingly complex array of adversarial attacks [33, 34]. This paper introduces a dual-layer defense mechanism that combines adversarial training with federated learning, providing robust protection against adversarial inputs while preserving user privacy. This method achieves a balanced attention to both performance and security, and mitigates vulnerabilities that can be easily missed by the conventional ones [35]

3.4 SCALING DATA PRIVACY SOLUTIONS

Data privacy remains a cornerstone of ethical AI development. Although differential privacy and federated learning methods are powerful [36], this paper points out their scalability issues in the area of global applications. Using synthetic data generation and privacy-preserving computation, we present solutions that reconcile privacy and performance and are such that AI systems are efficient and secure across borders [37, 38, 39, 40, 41].

3.5 ALIGNING POLICIES WITH LEGAL FRAMEWORKS

Standardization of privacy and security policy by jurisdiction is critical to align the development of AI with the legal regime [42]. This work highlights the need of proactive partnership among regulators, industry representatives, and scientists in order that flexible policies be created. Focusing on transparency and imbedding ethical considerations in AI algorithms, this paper argues for a less fragmented solution to fairness and accountability challenges in global AI governance [43].

4. CONCLUSION

The challenges associated with generative AI—such as openess and bias control, robustness against attacks and policy compliance—demonstrate a need for innovative and practical approaches. This paper makes new contributions towards addressing these constraints, e.g., an adaptive debiasing pipeline in which training data is flexibly conditioned to changes in demographic and cultural diversity over time. Not only does this methodology enhance fairness but also ensures that AI systems are fair and timely.

Specifically, the,adversarial training-based, dual-layered, defence strategy proposed, which integrates the two methods, provides an effective, anti-security attack prevention scheme, both for safeguarding, Generative AI, systems, and for any security risks to, the user against, security breaches, at the same time. These advances address key gaps in both observability and security, which set a foundation for the development of trustable/responsible AI systems. Although existing methods, e.g., differential privacy and federated learning, offer promising directions, their generalizability to scale and cross-border contexts pose a talking point. Conformity to legal and ethical frameworks through the anchoring of technical advancements to these frameworks is the foundation for this study in order to develop more responsible and more open AI practices. Solutions for these challenges require sustained collaboration among researchers, policymakers, and industry partners to design AI systems that are secure, fair, and agreeable to social norms.

REFERENCES

[1] Ashraf, I., Park, Y., Hur, S., Kim, S. W., Alroobaea, R., Zikria, Y. B., & Nosheen, S. (2023). A survey on cyber security threats in IoT-enabled maritime industry. IEEE Transactions on Intelligent Transportation Systems, 24, 2677–2690.

[2] Liao, Q., & Vaughan, J. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. ArXiv. https://arxiv.org/abs/2306.01941

[3] Mittermaier, M., Raza, M. M., & Kvedar, J. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. NPJ Digital Medicine, 6. https://doi.org/10.1038/s41746-023-00873-2

[4] Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of LLMs for illicit purposes: Threats, prevention measures, and vulnerabilities. ArXiv. https://arxiv.org/abs/2308.12833

[5] Sánchez, L., Grajeda, C., Baggili, I., & Hall, C. (2019). A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). Digital Investigation, 29(Supplement), S124–S142.

[6] Zhao, X., Ananth, P., Li, L., & Wang, Y.-X. (2023). Provable robust watermarking for AI-generated text. ArXiv. https://arxiv.org/abs/2306.17439

[7] Franzoni, V. (2023). From black box to glass box: Advancing transparency in artificial intelligence systems for ethical and trustworthy AI. Communication Systems and Applications, 118–130.

[8] Liao, Q., & Vaughan, J. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. ArXiv. https://arxiv.org/abs/2306.01941

[9] Hitron, T., Yaar, N. M., & Erel, H. (2023). Implications of AI bias in HRI: Risks (and opportunities) when interacting with a biased robot. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.

[10] Tang, N., Yang, C., Fan, J., & Cao, L. (2023). VerifAI: Verified generative AI. ArXiv. https://arxiv.org/abs/2307.02796

[11] Sethu, S., & Wang, D. (2022). A novel illumination condition varied image dataset-Food Vision Dataset (FVD) for fair and reliable consumer acceptability predictions from food. ArXiv. https://arxiv.org/abs/2209.06967

[12] Priestley, M. A., O'Donnell, F., & Simperl, E. (2023). A survey of data quality requirements that matter in ML development pipelines. ACM Journal of Data and Information Quality, 15, 1–39.

[13] Alotaibi, A. (2021). Demographic and cultural differences in the acceptance and pursuit of cosmetic surgery: A systematic literature review. Plastic and Reconstructive Surgery Global Open, 9. https://doi.org/10.1097/GOX.0000000000003456

[14] Virtanen, P., Gommers, R., Oliphant, T., Haberland, M., Reddy, T., Cournapeau, D., ... & Vázquez-Baeza, Y. (2019). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17, 261–272.

[15] Shi, Y., Liu, Z., Shi, Z., & Yu, H. (2023). Fairness-aware client selection for federated learning. In 2023 IEEE International Conference on Multimedia and Expo (ICME) (pp. 324–329).

[16] Gaudelli, N. M., Komor, A. C., Rees, H., Packer, M., Badran, A., Bryson, D., & Liu, D. R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature,

551, 457–463.

[17] Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. AI & Society, 1–12.

[18] Akter, S., Sultana, S., Mariani, M. M., Wamba, S., Spanaki, K., & Dwivedi, Y. K. (2023). Advancing algorithmic bias management capabilities in AI-driven marketing analytics research. Industrial Marketing Management.

[19] Xu, C., Zhang, J., Law, M., Zhao, X., Mak, P., & Martins, R. (2023). Transfer-path-based hardware-reuse strong PUF achieving modeling attack resilience with 200 million training CRPs. IEEE Transactions on Information Forensics and Security, 18, 2188–2203.

[20] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9, 211–407.

[21] Sebastian, G. (2023). Privacy and data protection in ChatGPT and other AI chatbots: Strategies for securing user information. SSRN Electronic Journal.

[22] Kaebnick, G., Magnus, D., Kao, A., Hosseini, M., Resnik, D. B., Dubljević, V., ... & Gordijn, B. (2023). Editors' statement on the responsible use of generative artificial intelligence technologies in scholarly journal publishing. Bioethics.

[23] Goldman, E. (2020). An introduction to the California Consumer Privacy Act (CCPA). SSRN Electronic Journal.

[24] World Medical Association Declaration of Helsinki. (2013). Ethical principles for medical research involving human subjects. JAMA.

[25] Johnson, J., Berg, T., Anderson, B., & Wright, B. (2022). Review of electric vehicle charger cybersecurity vulnerabilities, potential impacts, and defenses. Energies.

[26] Laux, J., Wachter, S., & Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. Regulation & Governance.

[27] Wallwey, C., & Kajfez, R. (2023). Quantitative research artifacts as qualitative data collection techniques in a mixed methods research study. Methods in Psychology.

[28] Hosseinzadeh, A., Eitel, A., & Jung, C. (2020). A systematic approach toward extracting technically enforceable policies from data usage control requirements. In International Conference on Information Systems Security and Privacy (pp. 397–405).

[29] Dingle, K., Camargo, C. Q., & Louis, A. (2018). Input–output maps are strongly biased towards simple outputs. Nature Communications, 9.

[30] Kang, M., Song, H., Park, S., Yoo, D., & Pereira, S. (2022). Benchmarking self-supervised learning on diverse pathology datasets. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3344–3354).

[31] Wang, P., Zhu, Y., Han, K., Yin, Z., Xiu, Q., & Hui, P. (2022). Fairness-aware algorithms for seed allocation in social advertising. In 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys) (pp. 464–473).

[32] Pain, C. D., Egan, G. F., & Chen, Z. (2022). Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement. European Journal of Nuclear Medicine and Molecular Imaging, 49, 3098–3118.

[33] Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning. ArXiv. https://arxiv.org/abs/1901.08573

[34] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. ArXiv. https://arxiv.org/abs/2307.15043

[35] Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., & Yan, S. (2023). Better diffusion models further improve adversarial training. In International Conference on Machine Learning. ArXiv. https://arxiv.org/abs/2302.04638

[36] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A., ... & Zhao, S. (2019). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14, 1–210.

[37] Semantha, F. H., et al. (2023). PbDinEHR: A novel privacy by design developed framework using distributed data storage and sharing for secure and scalable electronic health records management. Journal of Sensor and Actuator Networks, 12, 36.

[38] Fonseca, J., & Bação, F. (2023). Tabular and latent space synthetic data generation: A literature review. Journal of Big Data, 10, 1–37.

[39] Greenleaf, G. (2018). Global convergence of data privacy standards and laws: Speaking notes for the European Commission events on the launch of the General Data Protection Regulation (GDPR) in Brussels & New Delhi, 25 May 2018.

[40] Rouzrokh, P., et al. (2022). Mitigating bias in radiology machine learning: Data handling. Radiology: Artificial Intelligence, 4(5), e210290.

[41] Reer, A., Wiebe, A., Wang, X., & Rieger, J. (2023). FAIR human neuroscientific data sharing to advance AI driven research and applications: Legal frameworks and missing metadata standards. Frontiers in Genetics, 14.

[42] Duffourc, M., & Gerke, S. (2023). Generative AI in health care and liability risks for physicians and safety concerns for patients. Journal of the American Medical Association (JAMA).

[43] Phutane, A. S. (2023). Communication of uncertainty in AI regulations. Community Change.