Generative AI Disproportionately Harms Long Tail Users

DRAFT: This is not a camera-ready version yet

BARANI MAUNG MAUNG
Oxford Internet Institute, University of Oxford, UK, barani.maungmaung@gmail.com

KEEGAN MCBRIDE
Oxford Internet Institute, University of Oxford, UK, keegan.mcbride@oii.ox.ac.uk

JASON S. LUCAS
The Pennsylvania State University, USA, jsl5710@psu.edu

MARYAM TABAR
The University of Texas at San Antonio, USA, maryam.tabar@utsa.edu

DONGWON LEE
The Pennsylvania State University, USA, dongwon@psu.edu

ABSTRACT

Global discussions on Generative AI (GenAI) and its risks largely center around global hegemony in well-resourced regions. This article, instead, highlights how GenAI's risks can be further amplified in "long tail" populations and regions--those that speak low-resourced languages, are fragile, and marginalized. We recommend strategies to mitigate these risks and conclude by calling for a more nuanced and global dialogue on GenAI safety.

Additional Keywords and Phrases: Generative AI, Long Tail Users, Inclusive Technology, Global South

INTRODUCTION

Generative Artificial Intelligence (GenAI) is not new, but the recent release of tools such as OpenAI's ChatGPT/Sora, Google's Gemini, and Midjourney have made cutting-edge generative capabilities easily accessible to users with varying levels of technical skills. The democratization of access to such GenAI systems has enabled the generation of high-quality content in different mediums such as texts, codes, images, videos, or even their combinations with little to no cost.

The positive potential of GenAI is vast; for instance, it has been found to advance genetic research [1] and pharmacology [2]. Governments also stand to benefit as GenAI could aid in education and in planning public transport [3]. Despite the numerous benefits that may emerge from the widespread usage and availability of GenAI systems, several risks exist that must be considered. For instance, GenAI's ability to personalize content and develop realistic synthetic media (so-called deepfakes) may further exacerbate disinformation. Hallucinations (when GenAI models generate factually inaccurate information) and biased outputs (embedded into the technology through the model's training data) can have disastrous results if the generated insights are taken as truth. In addition, these biases can widen existing societal inequalities[1].

Moreover, despite GenAI being globally accessible, its societal impacts may not be evenly distributed worldwide. The World Health Organization (WHO) has warned that the use of medical GenAI systems in low-income countries could be “dangerous”. Similarly, we argue that GenAI’s impacts are highly influenced and shaped by geopolitical power structures, historical legacies, and the domestic institutional capabilities of the countries and regions in which it operates. This is pertinent, because the current global AI governance will remain largely insular as a result, favoring global hegemonies such as the United States, United Kingdom, and the European Union.

Big head vs. long tail countries

Thus, the benefits of GenAI are, to date, concentrated in countries with strong democracies that speak high-resource languages (those with large amounts of data for training GenAI, such as English). On the other hand, we posit that the risks of GenAI are greater in low-resource, fragile, and marginalized countries and populations.

Coined from the shape of distribution patterns, we use the term ‘big head’ to refer to countries that speak a high-resource language, have stable democracies, and have robust governance frameworks. Symmetrically, we use the term ‘long tail’ to indicate the opposite--low-resource, fragile, and marginalized countries--for instance, many countries from Africa, South or Southeast Asia, and the Caribbean and South America.

In big head countries, if an LLM provides inaccurate information about, or even a deepfake attack against, an individual, it may result in a defamation case. Yet, in long tail countries, the same situation may even result in the persecution of the individual. Similarly, false information on social media has instigated mass violence against individuals and communities in the past, leading to violent deaths in India and Nigeria. Thus, GenAI-based harms in long tail countries can have serious adverse outcomes.

Additionally, existing computational defenses against malicious usages of GenAI are mainly geared towards users in big head countries and predominantly trained with data in high-resource languages. Therefore, such defenses may not adequately handle the harms caused by GenAI in long tail countries and low-resource languages. For instance, research shows that “multilingual” tools that are built to detect AI-generated texts are 25% less effective in non-English languages [4]. Thus, it is time to recognize that the arsenal built for Western contexts may underperform outside of it.

Disinformation exacerbated

In long tail countries, where democracies are fragile and media systems tend to be harder to control, it is even more challenging to combat the proliferation of disinformation when GenAI is in use[2]. Trust is

[1] ProPublica. Machine Bias. (2016); https://bit.ly/4aDGNMu

[2] BBCNews. AI: Voice cloning tech emerges in Sudan civil war. (2023); https://www.bbc.co.uk/news/world-africa-66987869

integral for democracies to successfully operate, especially citizens' trust in public institutions and each other. With the democratization of GenAI, disinformation powered by persuasive and personalized machine-generated [5] texts and synthetic media can erode this trust. Though strong and stable democracies may combat this challenge with their rich media ecosystems[3], emerging and fragile democracies do not have such a luxury. These countries often lack the robust institutions (e.g., fact-checking and cybersecurity intelligence units), defense-technology (e.g., detecting AI models and technological infrastructure), and resources (e.g., financial and human expertise) necessary to counter the spread of false narratives effectively. Disinformation circulates rapidly undetected through less prominent channels, and in multiple languages. Malicious actors exploit existing social divisions and vulnerabilities by targeting specific linguistic, cultural, or social groups. This targeted approach makes it harder for authorities to identify and debunk false claims before they gain traction. In essence, the advent of GenAI has made the search for truth all the more challenging or even elusive, especially in regions where truth and transparency are needed the most.

Women and girls suffer

GenAI-powered disinformation may also disproportionately target long tail populations such as women and girls. In societies where women and girls are prized for their chastity, sexualized and gendered harassment is a prominent means of targeting them. In fact, it is already seen today, with women around the world being increasingly targeted with pornographic deepfakes[4]. In conservative, long tail communities where the rule of law remains weak, the impact of these attacks using GenAI could be much more severe. Victims will be socially ostracized and shamed, and lives will be upended. This will ultimately result in downstream implications for the representation and influence of women in politics.

Crises 2.0

As many forecasts, attacks using GenAI will be exacerbated during political crises and elections. Electoral periods can be incredibly sensitive, especially for countries with non-democratic institutions or a high risk of conflict [6], traits that often overlap with long tail regions. Existing social cleavages could widen and disinformation could make or break campaigns. During these times, GenAI has the potential to destabilize emerging democracies by allowing bad actors with few resources to efficiently create realistic but synthetic artifacts. This has already been illustrated in Bangladesh and Slovakia.

Fact-checking in these scenarios has been put forth as the solution. However, electoral times are characterized by high demand and supply for information; when large amounts of content are being uploaded every second, human fact-checking is not a scalable solution.

Big tech companies are taking preemptive automated initiatives to address AI threats, but their effectiveness remains uncertain, particularly in long tail settings. The COVID-19 infodemic highlighted the

[3] Harvard Kennedy School Misinformation Review. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. (2023); https://bit.ly/43I9JAH

[4] Context. Bollywood star or deepfake? AI floods social media in Asia. (2023); https://bit.ly/4cSkjJB

limitations of social platforms' countermeasures in combating the rapid spread of false information during a global crisis. Countermeasures developed in high-resource "normal" settings cannot be thoroughly tested under long tail circumstances, making it challenging to assess their efficacy in real-world scenarios, especially in events like upcoming elections. The nuanced and targeted nature of GenAI disinformation, spreading in multiple languages in various forms, further complicates the issue, making disinformation impacts challenging to anticipate, control, and address.

Data biases are not uniform

Finally, data bias embedded within GenAI systems poses yet another disproportionate risk for long tail users. In the West, gender and racial data biases are well-known phenomena. But in long tail countries, identity encompasses more than gender or race. For the African context, for example, scholar Chinasa Okolo has called for a nuanced non-Western understanding of identity, involving tribal affiliation and religion, to mitigate the risks posed by GenAI [7]. Similarly, in India, caste-biased outputs[5] by LLMs are already occurring.

However, auditing for biases for long tail contexts is a complicated endeavor for two reasons: it is almost impossible for developers to identify and remove such biases without context experts in place, and most long tail nations lack the governance frameworks and technical expertise to audit such biases [8]. Thus, long tail countries are subjected to slower response times from GenAI developers, making them more susceptible to adverse consequences of data bias.

Is regulation the answer?

Regulation has often been touted as a solution[6] for tackling the risks posed by GenAI. Yet, when discussing regulation, one must also consider whom these regulations will protect. Apart from a few exceptions (China, India, Brazil), the current global regulatory landscape on GenAI is largely situated in big head countries, including Europe and North America.

Policymakers within these countries will be, rightly so, developing policies that are relevant to their national contexts, that are unlike those in long tail countries. Additionally, even if long tail countries *do* implement regulation for GenAI, with large IT companies' revenues surpassing the annual GDPs of most of these countries, they lack the resources to effectively enforce regulations. Ultimately, a policy gap will remain, where regulations in big head countries will be incredibly impactful, but long tail countries will continue to be less protected from the risks of GenAI.

The question then arises, if regulation isn't a feasible solution for long tail countries, how can we tackle these issues?

Recommendations

[5] The Hindu. Racist, sexist, casteist: Is AI bad news for India?. (2023); https://bit.ly/43FB5Hq

[6] The New York Times. UN Officials Urge Regulation of Artificial Intelligence. (2023); https://nyti.ms/4cEnFzp

As a starting point, the onus for verifying information presented by GenAI technology should not rest solely on users, especially those in long-tail countries where it is increasingly difficult to do so due to a lack of digital literacy, reliable public data, and trust in public institutions.

Instead, academia, civil society, and companies producing GenAI must fill this gap. Academia and civil society should educate the public about the capabilities and limitations of GenAI, and occasionally verify critical news. In long tail countries where both academic and civil society landscapes lack robustness to shoulder this responsibility, it falls upon the companies creating GenAI technology to lead safety efforts for these long tail regions.

However, this direction may raise concerns about whose interests would be protected—industry, government, or society. Hence, companies should lead this initiative by involving multi-key stakeholders, especially those who can represent the interests of long tail users. Shaping safety innovations and policies through a multi-stakeholder approach can be promising, as it seeks to prevent any single entity from having undue influence or control over the process. Engaging citizens, government, academia, civil society organizations, and other stakeholders in development and policy can bring diverse perspectives and expertise together to inform inclusive policy decisions and solutions toward more representative, transparent, and accountable solutions.

GenAI models should be assessed internally and externally by diverse teams. Red teaming – the process of inviting external experts to assess GenAI models for risks – should not be limited to big head populations who reside in big head countries. To identify and address ongoing risks for long tail users, such feedback loops should be active in *all* stages of development.

Secondly, to combat GenAI-based disinformation more effectively, safety guardrails such as bias evaluations of training data, reducing graphic training data [9] and/or implementing policies to prohibit harmful content generation should be prioritized. If content policies are to be implemented, culturally specific policies are needed with context experts involved in this process [10], and all policies need to be continually updated to remain relevant.

Lastly, user-reporting systems of large GenAI companies should be made available in the languages of long tail countries. To illustrate, at the time of writing, OpenAI's ChatGPT and Google's Gemini feedback and reporting systems are available in one language – English. This is despite the fact that ChatGPT and Gemini are currently accessible in at least 30 non-English speaking countries, which can be classified as long-tail countries.

Conclusion

More recently, there have been developments towards a more international dialogue surrounding AI safety, extending from academia[7] to international diplomacy.[8] The authors recognize the significance of

[7] Far AI. Scientists Call for International Cooperation on AI Red Lines. (2024); https://far.ai/post/2024-03-idais-beijing/

[8] Foreign Policy. The U.N. Gets the World to Agree on AI Safety. (2024); https://foreignpolicy.com/2024/03/21/un-ai-regulation-vote-resolution-artifical-intelligence-human-rights/

these efforts in establishing a sustainable AI safety regime across regions. However, given that such endeavors are currently nascent and non-binding, the risks of GenAI in “long tail” countries persist. Therefore, we strongly advocate for a more comprehensive and inclusive global dialogue on GenAI safety across sectors, as the existing state remains largely insular, disproportionately focusing on affluent “big head” countries and their specific circumstances. Urgent action is imperative to rectify this imbalance before irreversible consequences and harms occur.

REFERENCES

[1] Cui, H. et al. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*. (Feb. 2024). DOI:https://doi.org/10.1038/s41592-024-02201-0.

[2] Chenthamarakshan, V. et al. 2023. Accelerating drug target inhibitor discovery with a deep generative foundation model. *Science Advances*. 9, 25 (Jun. 2023), eadg7865. DOI:https://doi.org/10.1126/sciadv.adg7865.

[3] Jittrapirom, P. et al. 2023. Visioning future transport systems with an integrated robust and generative framework. *Scientific Reports*. 13, 1 (Mar. 2023), 4316. DOI:https://doi.org/10.1038/s41598-023-30818-2.

[4] Macko, D. et al. 2023. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore, Dec. 2023), 9960–9987.

[5] Goldstein, J.A. et al. 2024. How persuasive is AI-generated propaganda? *PNAS Nexus*. 3, 2 (Feb. 2024), pgae034. DOI:https://doi.org/10.1093/pnasnexus/pgae034.

[6] Cazals, A. and Léon, F. 2023. Perception of political instability in election periods: Evidence from African firms. *Journal of Comparative Economics*. 51, 1 (Mar. 2023), 259–276. DOI:https://doi.org/10.1016/j.jce.2022.09.003.

[7] Okolo, C.T. 2023. The Promise and Perils of Generative AI: Case Studies in an African Context. *Proceedings of the 4th African Human Computer Interaction Conference* (East London South Africa, Nov. 2023), 266–270.

[8] Pashentsev, E. and Bazarkina, D. 2020. Malicious Use of Artificial Intelligence and International Psychological Security in Latin America. Report by the International Center for Social and Political Studies and Consulting. (Jun. 2020).

[9] Solaiman, I. et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. (Jun. 2023).

[10] Nkemelu, D. et al. 2023. Tackling Hate Speech in Low-resource Languages with Context Experts. *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development* (New York, NY, USA, 2023).