Algorithmic Gatekeepers:
The Human Rights Impacts of LLM Content Moderation

VI. Right to Non-Discrimination

European Center for
Not-for-Profit Law

Acknowlegements:

Author: Marlena Wisniak, ECNL.

We extend special thanks to Isabelle Anzabi, who contributed significantly during the early stages of the research process. We express gratitude to the ECNL team—Karolina Iwanska, Vanja Skoric, and Francesca Fanucci—for their thoughtful review and feedback on the report.

Valuable input and review were provided by Evani Radiya-Dixit from the American Civil Liberties Union (ACLU); Lindsey Andersen from Business for Social Responsibility (BSR); Mona Elswah and Aliya Bhatia from the Center for Democracy and Technology (CDT); independent researcher and policy expert Luca Belli; and Roya Pakzad from Taraaz.

Insightful contributions through interviews and consultations came from representatives of Meta's Human Rights Team, the Policy and Safety Machine Learning Teams at Discord, and the Research Team at Jigsaw.

We extend our sincere gratitude to everyone who generously contributed their invaluable time, insights, and expertise to the preparation of this report. Your thoughtfulness and creativity have greatly enriched the quality and depth of our findings. We thank Betsy Popken of the UC Berkeley Human Rights Center; Corynne McSherry from the Electronic Frontier Foundation (EFF); Daniel Leufer and Eliska Pirkova from Access Now; Dave Willner of Stanford University; Dunstan Alison Hope; Jonathan Stray from UC Berkeley; Justin Hendrix of Tech Policy Press; Mike Masnick of Techdirt; Paul Barrett from New York University; Sabina Nong of Stanford University; Tarunima Prabhakar from Tattle; and Vladimir Cortes.

Design for the publication was created by Sushruta Kokkula and Andrea Judit Tóth. The illustrations featured in the report are based on the work of Balázs Milánik, Rozalina Burkova (The Greats) and Daniela Yankova (The Greats).

We thank the Omidyar Network for their generous support.

April 2025

Table of contents

Applicability of international human rights law for AI governance 4

Right to Non-Discrimination 5

Applicability of international human rights law for AI governance

International human rights law, grounded in instruments like the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), and the International Covenant on Economic, Social and Cultural Rights (ICESCR), provides a tested and globally recognised framework for assessing the potential risks and benefits of AI systems and content moderation—and enables a right to remedy. Human rights principles recognise inalienable rights, such as privacy, non-discrimination, freedom of expression, and freedom of peaceful assembly and association, which must be protected from undue interference. While these protections were historically focused on government obligations, the UN Guiding Principles on Business and Human Rights (UNGPs) have established that businesses—including AI companies— also have a responsibility to respect and uphold human rights.

As AI governance frameworks proliferate, many companies rely on ethics-based or trust and safety-driven approaches to responsible AI. While these frameworks often emphasise fairness, accountability, and harm mitigation, they typically lack consistency, international legitimacy, and are voluntary. By contrast, a human rights-based approach, legally binding for States, offers a universal, internationally recognised, and adaptable framework that applies across jurisdictions and industries and provide a right to remedy.

Given that AI-driven content moderation impacts human rights, integrating these principles into AI development, use, and governance can help AI companies navigate trade-offs and mitigate harm. Ultimately, it will help them protect and promote human rights in their products, services, and activities. International human rights also serve as a common baseline that enables meaningful collaboration between AI developers, deployers, regulators, and civil society, making them an essential foundation for evaluating and addressing risks in generative AI and developing rights-respecting products.

This report aims to highlight the key human rights impacts of using LLMs for content moderation, with a focus on core civic freedoms. While it doesn't follow the methodology of a human rights impact assessment (HRIAs) under the UNGPs or a fundamental rights impact assessment (FRIAs) under the DSA or EU AI Act, our goal is to surface potential positive and negative impacts on a sector-wide level, to guide future HRIAs and FRIAs carried out by AI developers and deployers.

Right to Non-Discrimination

Legal basis

Article 2 ICCPR requires States "to respect and to ensure to all individuals [...] the rights [...] without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." This provides a robust legal framework to assess emerging technologies' impacts on marginalised groups, ensuring that they are not disproportionately impacted.

The UN Office of the High Commission for Human Rights (OHCHR) B-Tech warned that generative AI models can create harmful outputs targeting marginalised identities, amplifying false stereotypes, and facilitating widespread discrimination.[1] These models often overrepresent culturally dominant groups—such as white, Western, male, heterosexual, and cisgender individuals, often benefiting from colonial legacies—while misrepresenting or underrepresenting other groups at scale.[2] This imbalance entrenches harmful stereotypes, exacerbates existing biases and discrimination, and restricts marginalised communities' ability to control how their identities are portrayed in media and online spaces.[3]

As outlined in the U.S.NIST AI Risk Management Framework,[4] "Harmful bias in GAI systems can also lead to harms via disparities between how a model performs for different subgroups or languages (e.g., an LLM may perform less well for non-English languages or certain dialects). Such disparities can contribute to discriminatory decision-making or amplification of existing societal biases. In addition, GAI systems may be inappropriately trusted to perform similarly across all subgroups, which could leave the groups facing underperformance with worse outcomes than if no GAI system were used. Disparate or reduced performance for lower-resource languages also presents challenges to model adoption, inclusion, and accessibility, and may make preservation of endangered languages more difficult if GAI systems become embedded in everyday processes that would otherwise have been opportunities to use these languages."

1 OHCHR. (2023, November 2). Taxonomy of human rights risks connected to generative AI: Supplement to B-Tech's foundational paper on the responsible development and deployment of generative AI. https://www.ohchr.org/en/documents/tools-and-resources/taxonomy-generative-ai-human-rights-harms-b-tech-gen-ai-project

2 Ibid.

3 Office of the United Nations High Commissioner for Human Rights. (n.d.). B-Tech: Taxonomy of GenAI human rights harms – Right to non-discrimination. Retrieved from https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf.

4 U.S. National Institute of Standards and Technology (NIST). (2024, July). Artificial Intelligence Risk Management Framework (p. 8). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Bias in datasets

LLMs may potentially perform better than the current set of content moderation systems and processes, which are known for biases and inconsistency due to human moderators making variable decisions, among other reasons. That said, LLMs (like machine learning systems) also have systemic biases, not only because of data they are trained on, but also due to broader root issues. Indeed, the erasure and marginalisation of non-dominant groups by colonial powers has resulted in the dominance of cultural narratives that shape online texts, which are subsequently used as training data. These biases, influenced by both the data and the contexts in which models are used, significantly impact marginalised communities and how they're represented in society. While efforts are being made to reduce bias, such as improving model performance through metrics like the true positive rate (e.g., optimising the ROC curve[5]), deeper issues remain. Systemic inequalities in training data and labeling processes continue to create challenges that are not fully addressed.

One key issue is that generative AI models often reproduce harmful or derogatory views over-represented in datasets the models are trained on, views that are particularly biased against marginalised groups. This includes amplifying false and harmful stereotypes, perpetuating discrimination, and fostering inequality at scale. For example, outputs frequently overrepresent dominant cultural groups—such as white, Western, male, heterosexual, and cisgender individuals—while underrepresenting or misrepresenting others. This disparity entrenches stereotypes and limits the ability of marginalised groups to control their representation in media and digital spaces.

LLMs may systematically (quickly and at a large scale) discriminate against marginalised groups online. Generative AI has an established record of bias based on race,[6] gender,[7] and religion.[8] Thus, LLMs used for content moderation will likely lead to both direct and indirect discrimination. For example, if a system has only been trained on instances of harassment involving a man and a woman, it may fail to recognise cases of harassment within gender-diverse or LGBTQIA+ communities.

Indeed, bias within these systems exacerbates discriminatory outcomes, as seen in the over-moderation of content of marginalised groups and the under-moderation of dominant cultural narratives (see above section on freedom of expression). Content moderation may appear "neutral on its face but hits disproportionately at particular groups, and does so without any objective justification."[9] The clear bias within training data and preliminary testing means that any differentiation of treatment is neither

---

5 Wikipedia contributors. (n.d.). Receiver operating characteristic. Wikipedia. Retrieved April 2, 2025, from https://en.wikipedia.org/wiki/Receiver_operating_characteristic
The Receiver Operating Characteristic (ROC) curve is a graphical tool used to assess the performance of binary classification models by plotting the true positive rate (sensitivity) against the false positive rate across various threshold settings. This visualization illustrates the trade-offs between correctly identifying positive instances and incorrectly classifying negative ones. The area under the ROC curve (AUC) serves as a summary metric, with higher values indicating superior model performance in distinguishing between classes.

6 Schreiber, M. (2024, March 25). Why large language models like ChatGPT treat Black and White sounding names differently. Stanford HAI. https://hai.stanford.edu/news/why-large-language-models-chatgpt-treat-black-and-white-sounding-names-differently

7 UNESCO. (2024, March 27). Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes. UNESCO. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

8 Toushik Wasi, A., Islam, R., Islam, M. R., Rafi, T. H., & Chae, D. K. (2024, July 25). Measuring the Impact of Large Language Models on Underrepresented Groups. arXiv. https://arxiv.org/html/2407.18376v1

9 Lester, L. of Herne Hill, Q. C. (n.d.). Non-discrimination in International Human Rights Law (p. 12). The Commonwealth iLibrary. Retrieved from https://www.thecommonwealth-ilibrary.org/index.php/comsec/catalog/download/439/439/3813?inline=1

reasonable nor objective.[10] Even if the aim of LLM integration into content moderation is to achieve a legitimate purpose, the clear differentiation of treatment is unacceptable. Many generative AI systems are trained on large-scale datasets derived from the open web, such as CommonCrawl dumps.[11] These datasets reflect societal biases, including racial, cultural, and religious prejudices, as well as toxic content and illegal behaviors.[12] Moreover, an expanding body of research reveals that bias in multimodal models—those that process combinations of text, image, audio, and video inputs—not only inherit societal and historical stereotypes but often amplify them, sometimes to a greater extent than models limited to a single modality.[13]

Research has shown that as the scale of these datasets increases, so too does the amplification of harmful biases.[14] For example, as the scale of the model increased, the errors in associating images of human faces with offensive classifications, rather than the "human being" class, increased by 50%. Alarmingly, this scaling effect exacerbated harmful biases: the likelihood of associating Black female faces with the "criminal" class doubled, and for Black male faces, this association increased fivefold.[15]

The limitations of "debiasing"

Efforts to address these biases face significant challenges. One approach involves using pre-cleaned, regulation-compliant datasets or synthetic data to retrain and fine-tune LLMs.[16] However, these methods remain limited by the lack of value pluralism in the labeling processes, among other issues. While labeling is increasingly outsourced to (underpaid) workers in the Global Majority,[17] practices and guidelines are often centralised within culturally homogenous groups, such as those in Silicon Valley, whose values may not reflect the diverse perspectives needed for global applicability.

A clear conception of what the biases are, how protected groups should be classified, and who should be included in these groups, is needed in each language and country. No one expects platforms have this or can ever achieve this.[18] Zooming out, it's important to note that debiasing methods do not eliminate the discriminatory effects of AI systems. These approaches place decision-making power in the hands of service providers rather than policymakers and affected communities, allowing them to define what qualifies as discrimination, determine when it occurs, and choose how to address it.[19]

Research has further exposed the challenges in understanding and evaluating the biases of LLMs in recommender systems, for example, noting that "in terms of intrinsic fairness, which does not involve direct sensitivity, unfairness across demographic

10 Ibid., p.14
11 Birhane, A., Prabhu, V., Han, S., & Boddeti, V. N. (2023, June 28). On hate scaling laws for data-swamps. (p.1) arXiv. https://arxiv.org/abs/2306.13141
12 Atlantic Council. (2023). Scaling Trust on the Web: Comprehensive Report. Atlantic Council. Retrieved from https://www.atlanticcouncil.org/wp-content/uploads/2023/06/scaling-trust-on-the-web_comprehensive-report.pdf
13 The Dark Side of Dataset Scaling. P 1230
14 Birhane, A., Prabhu, V., Han, S., & Boddeti, V. N. (2023, June 28). On hate scaling laws for data-swamps. arXiv. https://arxiv.org/abs/2306.13141
15 Ibid.,
16 Atlantic Council. (2023). Scaling Trust on the Web: Comprehensive Report. Atlantic Council. Retrieved from https://www.atlanticcouncil.org/wp-content/uploads/2023/06/scaling-trust-on-the-web_comprehensive-report.pdf
17 Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. TIME. Retrieved from https://time.com/6247678/openai-chatgpt-kenya-workers/.
18 Jonathan Stray, personal communication, August 23, 2024.
19 EDRi. (2021, September 21). If AI is the problem, is debiasing the solution? European Digital Rights (EDRi). https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/

groups remains significant." What's more, many scholars have raised concerns about the potential harms of uncritically defining demographic traits such as gender, race, and disability. A key issue is that using demographic data in this way risks reinforcing and essentialising socially constructed categories.[20]

Finally, while the use of synthetic data could help address the lack of diversity in training datasets, the limitations of current labeling processes and the absence of input from external stakeholders restrict the effectiveness of such efforts.[21] More importantly, synthetic data can never adequately replace real data reflecting marginalised' groups views and the use of such data could be in tension with these groups' right to human dignity.

Multilingual models and the "resourcedness" gap

Multilingual language models (MLLMs) are a commonly proposed solution to previous assertions of poor and inconsistent content moderation across languages. Yet, these models face significant shortcomings, particularly in content analysis and equitable performance across languages. These challenges arise from limitations in training data and structural inequalities between high- and low-resource languages.
MLLMs are often trained on machine-translated text because of the paucity in available data in languages other than English.[22] This issue reflects differences in the availability, quality, and diversity of training data for various languages. Both historically and in the present, imperial powers have systematically erased non-English languages, particularly regional and indigenous ones.

Consequently, the vast majority of online content available for training, tuning, and testing language models is in English.[23] English is the most well-resourced language by a significant margin, followed by languages like Spanish, Chinese, and German, which also have sufficient high-quality datasets to support language model development.[24] Medium-resource languages, such as Russian, Hebrew, and Vietnamese, have fewer but still robust datasets, while low-resource languages, such as Amharic, Cherokee, and Haitian Creole, lack the necessary volume and quality of training data to build effective language models.[25] Researchers found that training data for low-resource languages is often misclassified, mistranslated, or derived from narrow domains, such as religious texts or Wikipedia, making it disconnected from how people actually speak.[26]

A critical limitation in content analysis tasks is their inability to perform consistently well across all languages. As highlighted in the Center for Democracy and Technology's (CDT) report "Lost in Translation – Large Language Models in Non-English Content

20 Yee, K., Redfield, O., Sheng, E., Eck, M., Schoenauer, A., & Belli, L. (2022, October). A keyword based approach to understanding the over penalization of marginalized groups by English marginal abuse models on Twitter. arXiv. https://arxiv.org/abs/2210.06351
21 Sabina Nong, personal communication, September 13, 2024.
22 Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 560-570. https://aclanthology.org/2020.acl-main.560/
23 Ibid.
24 Ibid.
25 Nicholas, G., & Bhatia, A. (2023, May 23). Lost in translation: Large language models in non-English content analysis. Center for Democracy & Technology. (p.19). https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/
26 Ibid., p. 6.

Analysis,"[27] this challenge arises from the "curse of multilinguality."[28] This concept, developed by researchers Conneau and Khandelwal in 2020, posits that LLM developers must balance performance across languages, especially given the scarcity of high-quality language data.[29] As a result, companies may prioritise languages spoken in wealthier regions or their primary markets, leaving marginalised communities further behind. This tradeoff perpetuates inequalities by deprioritising languages spoken by less politically or economically influential groups.[30]

The implications of these disparities are profound. Underrepresentation of low-resource languages in generative AI training datasets leads to the underperformance of these models for speakers of such languages. Errors in machine translation undermine the reliability of models and contribute to performance inconsistencies. Additionally, when MLLMs fail, their issues are often opaque and difficult to diagnose due to the complex and unintuitive connections they make across languages.[31] As cautioned by the UN OHCHR B-Tech, this underperformance can itself constitute a form of discrimination, widening the digital divide between high- and low-resource regions. Moreover, the concentration of generative AI development in the Global North accelerates data production and usage in these regions, while deepening "data poverty" in the Global Majority. This lack of access to data can negatively impact economic development and undermine human rights in underrepresented regions.[32]

Failure to understand cultural and societal context

As noted in "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," LLMs encode biases, and these biases extend beyond protected attributes to encompass identity characteristics that are deeply rooted in culture. Toxicity classifiers, for example, require culturally appropriate training data for effective auditing in different contexts. However, even with such data, marginalised identities may still be overlooked if there is no understanding of what to audit for.[33]

Language models also struggle to account for changes in language over time, such as evolving slang or shifts in usage, and their reflections of language across different contexts are similarly limited. LLMs often struggle with providing up-to-date information because their training data is fixed at a certain point in time. This makes them less useful for contexts where current knowledge is essential. One approach to mitigating this problem is using real-time retrieval methods that enable models to pull the latest information from external sources—e.g., databases, news sources, or APIs—to access and integrate fresh data during inference.

27 Ibid.

28 Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020, April). Unsupervised cross-lingual representation learning at scale. arXiv. https://arxiv.org/pdf/1911.02116

29 Ibid.

30 Nicholas, G., & Bhatia, A. (2023, May 23). Lost in translation: Large language models in non-English content analysis. Center for Democracy & Technology. (p.28). https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/

31 Nicholas, G., & Bhatia, A. (2023, May). Lost in Translation: Large Language Models in Non-English Content Analysis (p. 30). Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf.

32 OHCHR. (2023, November 2). B-Tech: Taxonomy of GenAI and Human Rights Harms. Office of the High Commissioner for Human Rights. Retrieved from https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf.

33 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Another challenge for LLMs is keeping up with how language changes over time, such as the emergence of new slang, shifts in word usage, or changes in cultural context.[34] Because training data is typically static, models can't adapt to these changes without updates or fine-tuning. These models operate by transferring knowledge between language contexts, but this frequently results in higher-resource languages overwriting the contexts of lower-resource ones.

This issue is compounded by the reliance on translated text or limited sources such as Wikipedia and the Bible, as explained above, rather than native speakers' language usage. Consequently, tasks requiring local contextual understanding, such as hate speech detection, often yield poor outcomes in low-resource languages.[35] A significant example of this is in African languages, which account for one-third of the world's languages. Many of these are oral languages that are slowly disappearing as native speaker populations decline. LLMs developed by Western-based tech companies fail to adequately serve these languages, as they do not account for the cultural and contextual relevance specific to local speakers.[36]

Community-driven solutions

Building AI systems is prohibitively expensive and restrictive. Efforts to develop AI tools tailored to local and regional contexts in the Global Majority face significant challenges namely due to inadequate funding and poor infrastructure, colored by colonial and imperialist dynamics. In Africa, for example, research to provide training data in non-dominant languages is hindered by underfunded linguistics departments, the declining use of native languages, and limited machine-readable data. Additionally, issues like insufficient internet access and a lack of domestic data centers restrict developers' ability to deploy advanced AI capabilities.[37]

Several initiatives aim to address these gaps. In Nigeria, Awarri, a government-endorsed AI startup, is working to create the country's first LLM to integrate Nigerian languages into AI tools.[38] Similarly, Masakhane, an organisation promoting natural language processing (NLP) for African languages, has released over 400 open-source models and 20 African-language datasets since its founding in 2018.[39] EqualyzAI, a startup, seeks to preserve African languages through digital tools, developing voice tools and AI models covering 517 African languages.[40]

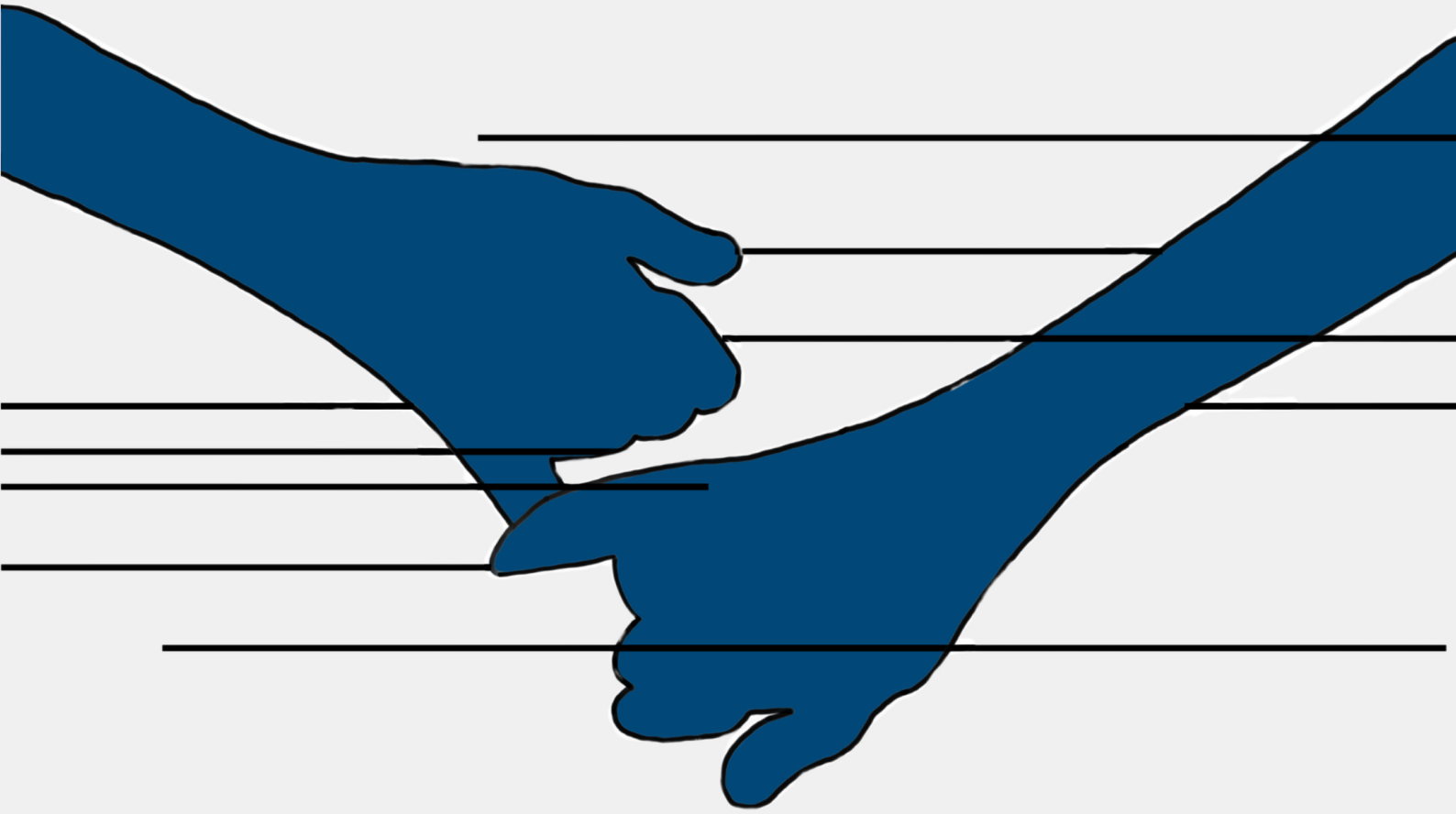
One notable milestone is Lelapa AI's release of InkubaLM, a small language model supporting IsiXhosa, Yoruba, Swahili, IsiZulu, and Hausa.[41] Trained on two open-source datasets with 1.9 billion tokens, InkubaLM was developed through community-driven efforts, including workshops where native speakers created data for the model. Despite its smaller scale, InkubaLM performs comparably to larger models on tasks like

34 Elswah, M. (2024, September 27). Moderating Maghrebi Arabic content on social media. Center for Democracy & Technology. https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/
35 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). ACM. https://doi.org/10.1145/3442188.3445922
36 Tsanni, A. (2024, November 11). What Africa needs to do to become a major AI player. MIT Technology Review. https://www.technologyreview.com/2024/11/11/1106762/africa-ai-barriers/
37 Ibid.
38 Ibid.
39 Ibid.
40 Ibid.
41 Lambebo Tonja, A., Dossou, B. F. P., Ojo, J., Rajab, J., Thior, F., Wairagala, E. P., Anuoluwapo, A., Moiloa, P., Abbott, J., Marivate, V., & Rosman, B. (2024, August 30). InkubaLM: A small language model for low-resource African languages. arXiv. https://arxiv.org/html/2408.17024v1

translation and sentiment analysis.[42] This community-informed approach, guided by linguistic and cultural representatives, is much more inclusive and effective.[43]

Organisations like Lelapa AI continue to build datasets and NLP tools for African languages, addressing language-specific challenges through collaborations with local developers and researchers. As noted by CDT, these efforts align with global movements to develop NLP tools for underrepresented languages, focusing on Arabic, Indian, African, Indonesian, and indigenous American languages.[44]

Finally, LLMs could present opportunities to include community-driven counter-speech narratives in moderation and safety feature responses. For example, counter-speech initiatives, such as those addressing anti-Muslim stereotypes, can leverage LLMs to create humor-based content to debunk hateful information. A project like the "#MuslimsReportingStuff" campaign humorously challenged stereotypes, demonstrating the potential of AI to facilitate more effective counter-narratives.[45] While counter-narratives have their own challenges, which are out of scope for this report, this application could open new avenues for combating harmful biases and stereotypes through innovative and culturally nuanced AI tools.



42 Tsanni, A. (2024, November 11). What Africa needs to do to become a major AI player. MIT Technology Review. https://www.technologyreview.com/2024/11/11/1106762/africa-ai-barriers/
43 Ibid.
44 Radiya-Dixit, E., & Bogen, M. (2024, October). Beyond English-centric AI: Lessons on community participation from non-English NLP groups (p. 2). Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2024/10/2024-10-18-AI-Gov-Lab-Beyond-English-Centric-AI-brief-final.pdf
45 Roya Pakzad, personal communication, September 5, 2024.

Challenges of evaluation

Evaluating AI systems, particularly LLMs, remains extremely immature, especially when considering multilingual content and content that is created by or affects marginalised groups. Many benchmarks used to assess "AI safety" and fairness are direct translations of English benchmarks, which fail to capture linguistic and cultural nuances in diverse languages. Safety assessments for racialised groups, women and non-binary persons, LGBTQIA+ communities, and other marginalised groups, are scarce. This also applies to low- and medium-resource languages, although it's encouraging to see community-driven evaluation efforts emerge. For example, the IndoNLP group developed NusaX, a human-translated benchmark dataset covering 10 low-resource Indonesian languages.[46] Yet the overall lack of specific expertise in these areas results in significant gaps in evaluation, leaving key risks unaddressed.[47]

Indeed, the disparity in generative AI capabilities across languages highlights the urgent need for high-quality multilingual training datasets and benchmarking frameworks. Without robust evaluation mechanisms, AI products and services with a global user base, including LLM moderation, will continue to perform inconsistently across different linguistic and cultural contexts. As noted in the DTSP report, addressing this gap is crucial for ensuring fairness and accuracy in AI systems.[48]

Additionally, benchmarks to assess the authenticity of content are often rooted in Western-centric norms, leading to biased outcomes. In the West, factors such as mononyms, generic profile pictures, repetitive content uploads, and high follower counts may be flagged as indicators of fake accounts.[49] However, in many parts of the Global Majority, individuals commonly use mononyms or pseudonyms for cultural reasons or to maintain anonymity in politically sensitive environments.[50] AI classifiers, trained primarily on Western data, frequently misinterpret these accounts as fake, leading to unjust account suspensions and disproportionate harm to already marginalised users.[51]

Addressing these evaluation challenges requires AI developers and deployers fine-tuning these models to invest in culturally diverse datasets, multilingual benchmarking, and context-aware evaluation frameworks to ensure that AI systems serve all users equitably.

---

46 Radiya-Dixit, E., & Bogen, M. (October, 2024). Beyond English-centric AI: Expanding global inclusion in AI governance. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2024/10/2024-10-18-AI-Gov-Lab-Beyond-English-Centric-AI-brief-final.pdf

47 Roya Pakzad, personal communication, September 5, 2024.

48 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/;
Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S. (2023, October 22). MEGA: Multilingual evaluation of generative AI (Version 4). arXiv. https://arxiv.org/abs/2303.12528

49 Maung Maung, B. (2023). When conflict goes online: How trust & safety systems fall short in handling crises in the global majority. Tech Global Institute. https://techglobalinstitute.com/research/when-conflict-goes-online-how-trust-safety-systems-fall-short-in-handling-crises-in-the-global-majority/.

50 Ibid.

51 Ibid.

Algorithmic Gatekeepers: VI. Right to Non-Discrimination

European Center for
Not-for-Profit Law