Venkat Rao ***is a seasoned technology leader with extensive experience in guiding large-scale digital transformation initiatives and integrating AI technologies. A strong advocate for inclusive design and accessibility, Venkat has consistently championed the creation of accessible and empowering technology solutions. He is also the Founder and Author of the Assistive Technology Blog (https://assistivetechnologyblog.com), a digital publication dedicated to showcasing innovations that empower individuals with disabilities.***

Dr. Emily Springer ***is a responsible AI consultant and an AI literacy trainer for non-coders through The Inclusive AI Lab. She is a member of UNESCO's Women for Ethical AI, a certified algorithmic auditor, and has been recognized as Top 100 Brilliant Women in AI Ethics. Drawing on her experience in international development and consulting for UNFPA, Data2X, World Bank, and others, Dr. Springer is committed to ensuring AI works for all people, especially the global majority, and trains practitioners and organizations to understand and implement inclusive AI.***

ISSN :2582-8304

Building AI Systems That Recognize Global Diversity through better data, smarter tech, and inclusive practices

Venkat Rao

Dr. Emily Springer

venkat@accessibleintelligence.solutions

Abstract

While Large Language Models (LLMs) like ChatGPT have been heralded as revolutionary technologies since their mainstream debut in late 2022, their design, data, and deployment reflect a narrow, Western-centric perspective that fails to account for global diversity. This article explores the cultural and ethical implications of AI systems trained predominantly on data from WEIRD (Western, Educated, Industrialized, Rich, Democratic) societies, emphasizing how such bias undermines the inclusivity and effectiveness of these tools for non-Western users. Through real-world examples—from AI writing assistants erasing cultural nuance to facial recognition misjudging emotions across cultures—the article illustrates how AI can perpetuate harm when it "gets culture wrong." It argues for a transformative approach to AI development rooted in global equity, recommending strategies such as inclusive data practices, culturally adaptive technologies, and participatory design processes. The goal is to reimagine AI as a tool that connects rather than divides, ensuring it reflects and respects the full spectrum of human culture and language in our increasingly AI-driven future.

Keywords

Cultural Bias in AI, Inclusive Artificial Intelligence, Global AI Ethics, Large Language Models (LLMs), Data Diversity, Cross-Cultural Technology Design, Equitable AI Development

Building AI Systems That Recognize Global Diversity through better data, smarter tech, and inclusive practices

Introduction

Although many claim that "the world" was introduced to a new technology—Large Language Models (LLMs) like ChatGPT—in late 2022, this statement fails to capture the nuances and challenges of making technologies effective and meaningful for people living in different cultures, speaking different languages, and having differential access to digital devices. In order to achieve AI products that work well across global differences, we need to reimagine how we build and use AI across cultures by improving datasets, building smarter tech, and implementing inclusive practices.

Since late 2022, those with access to digital devices and the internet have been introduced to a completely new way of getting answers to practically any question—by interacting with a type of AI called a Large Language Model (LLM). Examples include ChatGPT, Claude, Gemini, DeepSeek, and Perplexity. What makes these tools novel is their ability to *generate* new text, taking a user's input (a question), processing a likely answer, and returning a computed text answer. As corporations and governments drive to capture economic value from generative AI, people and communities interested in inclusion should be asking: For whom do these tools work well and why? How can we adopt new technologies safely, responsibly, and inclusively? How can

we, as citizens, play a role in building AI that ethically allocates resources and opportunities in our new AI-driven future?

But before we can assess how well Large Language Models work for non-Western audiences (i.e., the global majority!), we need to understand a bit more about how LLMs are typically built. It is important for us to also understand how all of the LLMs work under the hood, especially for global, non-Western audiences. We may not explicitly realize or even notice it, but where we are located, who we identify as, the societies we belong to, and the cultures we grow up in shape how we see the world, how we talk, and what we value. The problem is, the AI often doesn't get this.

Before deployment, LLMs go through a process of "training." This means that LLMs are provided access to a massive amount of information, often combinations of text-based datasets (like Wikipedia) and internet scrapes. One of the most popular data sources is the Common Crawl, which is created by scraping publicly available information on the internet. Datasets like these are then used to "train" LLMs to understand the semantic relationship between words in sentences, paragraphs, and entire books. After this process, a well-trained LLM will be able to respond to a user's natural question with a response that, although generated through computation, feels to the user like it was constructed by a human who understood their question.

But wait! Most datasets are by, for, and about Western audiences—and those audiences are just plain WEIRD – Western, Educated, Industrialized, Rich, and Democratic. They just don't represent a

global audience. In addition, when considering these models for a global audience, there are two key issues to understand and follow:

1. **Serious debates are occurring around the world with respect to the legality of training generative AI on copyrighted information, even if accessible via the internet. Consider the recent case of OpenAI releasing the Studio Ghibli GPT. What makes this example striking is that a Western corporation is gaining users by training on the copyrighted work of famed animator and filmmaker Hayao Miyazaki (cofounder of Studio Ghibli, an animation studio in Japan), which is antithetical to his entire approach to art and technology.**
2. **LLMs are computational machines, and they do not understand, think, feel, etc., no matter how "real" the answer may feel to the user. Dominant understandings of factually incorrect output information are that those responses are "hallucinations"-abnormalities rather than a core product feature of generative AI. A more useful framing is to view incorrect information as demonstrative of poor training data, poor product tooling, among others. When taken globally, LLMs are more likely to hallucinate concerning non-Western cultures because of their training data.**

When AI Gets Culture Wrong: The Harm of Western Bias

Think about the massive amount of information these LLMs learn from. Most of it comes from the Western world, primarily written in English. This means the AI learns mostly about Western culture, Western ways of speaking, and Western values. It develops a kind of "Western viewpoint" by default. So, what happens when this AI

interacts with someone from India, Africa, or East Asia? It often leads to misunderstandings, poor results, and sometimes, real harm.

Let's look at some real-world examples where this Western bias causes problems:

- **Losing Your Voice to AI Writing Assistants: AI tools offer suggestions to improve writing. Sounds helpful, right? But if the AI is trained mainly on Western writing styles, its suggestions often push users toward that style. A study found that Indian participants started writing more like Americans when using AI help, even describing their food and festivals in a way an outsider would. *The harm?* This isn't just about grammar; it's about potentially losing the unique flavor and nuances of one's cultural expression. AI risks making us all sound the same, erasing the richness that cultural diversity brings to communication.**

- **The "Angry" African Child & the "Untrustworthy" Japanese Colleague: Emotion AI tools try to read feelings from facial expressions. But they're usually trained on Western faces. In the West, a big smile often means happiness, and maybe less expression means something negative. But that's not universal! In parts of Southeast Asia, a smile might hide embarrassment. In Japan, showing less emotion is often seen as respectful and professional, but an AI might wrongly flag this as "unemotional" or even "untrustworthy". An African child showing a neutral expression, unfamiliar to the AI, was even flagged as "angry". *The harm here?* When emotion AI tools are integrated into government services, company hiring processes, and more, people can be unfairly judged, lose job opportunities, or face**

suspicion simply because their culturally normal expressions don't match the AI's Western training. Imagine being detained at an airport because facial recognition, biased by Western norms, misread your neutral expression as aggression, as has happened to Arab travelers.

- **Translation Troubles : More than just words : AI translation is getting better, but it struggles deeply with cultural context. Slang is a great example. Meta's AI translated the Spanish slang "no manches" (meaning "no way !" or "come on !") literally as "no stain," which is meaningless. AI also misses vital distinctions, like in the Yoruba language, where there are different words for "queen" and "wife of the king." AI often translates both simply as "queen," erasing an important cultural difference. *The harm ?* At best, it leads to confusing or awkward communication. At worst, it can cause serious misunderstandings, strip away important cultural meanings, and make the technology frustrating or unusable for non-Western users. An Afro-Indigenous Brazilian man's asylum application was even delayed because of errors in AI-powered communication that couldn't handle his dialect.**

- **Seeing the World Through Western Eyes: Computer vision AI, which identifies things in images, also suffers. Models trained on Western images perform poorly when shown culturally diverse scenes. They might misidentify traditional clothing or objects. Text-to-image generators often portray non-Western cultures based on stereotypes or an "outsider's" romanticized view rather than reality. *The harm?* This reinforces stereotypes, spreads**

inaccurate representations of cultures, and makes AI less useful for tasks involving non-Western visual contexts.

These aren't just minor glitches. They show that an AI trained primarily in one culture *cannot* effectively or fairly serve everyone else.

How Do We Fix This? Building AI that Works for the World

If the problem starts with the data and the design being too Western-focused, the solution must involve broadening that focus significantly. We need to actively build AI that is able to analyze and process globally diverse inputs. Here's how we can add more substance to make that happen:

1. **Get Better, More Diverse Data: This is job number one. Simply pouring more Western data into AI won't fix it. We need *representative* data.**
    - **Go Global: Actively collect text, speech, and images from Asia, Africa, Latin America, Indigenous communities, and other underrepresented regions. This includes supporting efforts to digitize languages that have few online resources.**
    - **Be intersectional: Individuals aren't just "man" or "woman"; they are dynamic beings that have age, sexualities, education, parental status, religions, and more. Representative, intersectional datasets are needed to represent the complexity of each person. Community-Centric Approach: Don't just take data; work *with* communities. Involve local people in collecting and labeling information. They understand the nuances best and can**

prevent misinterpretations. This ensures the data truly reflects the culture.

- **Expert Collaboration: Bring in local linguists, sociologists, and cultural experts. Their insights are invaluable for designing data collection that is culturally sensitive and accurate.**
- **Culturally Aware Labeling: How data is labeled matters. Guidelines for annotating data need to account for cultural context to avoid baking in biases due to labeling.**

2. Smarter Tech Approaches: We also need technical innovations designed for cultural adaptability.

- **Teaching AI to Transfer Knowledge: Techniques like *transfer learning* allow AI trained on lots of data (like English) to learn a new, low-resource language or cultural context much faster, even with less data. *Few-shot learning* helps AI learn new concepts (like a specific cultural greeting) from just a few examples. This helps address existing data gaps.**
- **Understanding Language Structure: For languages with non-Latin scripts (like Hindi or Arabic) or complex grammar (like Turkish or Basque), AI needs specialized techniques. This includes better *tokenization* (how AI breaks down words) that respects the language's structure and using phonetics (like the International Phonetic Alphabet) to help AI understand script differences.**
- **Detecting and Reducing Bias: We need tools specifically designed to find and mitigate cultural biases in both the data**

and the AI models themselves. This is an ongoing process, not a one-time fix.

- **Building Context Awareness: AI needs to get better at understanding context. This involves techniques to help AI grasp implied meanings, non-verbal cues (though this is hard!), and the overall situation, not just the words themselves.**

3. Inclusive Design and Development: Who builds the AI and how they build it matters.

- **Diverse Teams: If the people building AI come from varied cultural backgrounds, they are more likely to spot potential issues and build more inclusive systems. We need more diversity in the AI field globally.**
- **Ethical Frameworks and Governance: Companies and countries need clear guidelines and rules. International standards like the UNESCO Recommendation on AI Ethics emphasize protecting cultural diversity. We need governance models that involve multiple stakeholders (governments, companies, and communities) to ensure that cultural representation is prioritized. This includes audits to check for bias.**
- **Community Involvement in Design: Beyond just data collection, involve communities in designing and testing the AI (participatory design). Does it meet their needs? Is it culturally appropriate? Continuous feedback loops are essential.**

Putting these pieces together – better data, smarter tech, and inclusive practices – is how we move toward AI that genuinely understands and respects global cultures.

Looking Ahead: AI That Connects, Not Divides

Imagine an AI that doesn't just translate words but understands the cultural meaning behind them: An AI that can adapt its communication style, recognize different cultural norms, and avoid both overt and covert stereotypes. This kind of AI could be a powerful tool for connection and ensure that the economic value of AI does not merely reproduce existing global inequalities. It could help people from different backgrounds understand each other better, preserve endangered languages and traditions, and provide truly personalized and respectful assistance to everyone, everywhere.

Getting there requires a real shift in how we build AI. We need to move away from the current model where Western data dominates and actively embrace global diversity at every stage – from data collection to team building to the final product. It's about making a conscious choice to build AI that respects and reflects the incredible variety of human culture. Only then can AI fulfill its potential as an inclusive technology for the 21st century.

References

- *AI Competence. (n.d.). Emotional AI Fails Globally: Western Bias Exposed. https://aicompetence.org/emotional-ai-fails-globally-western-bias-exposed/*
- *Agarwal, V. et al. (2024). No Filter: Cultural and Socioeconomic Diversity in Contrastive Vision-Language Models. arXiv. https://arxiv.org/pdf/2405.13777*
- *Al Kuwari, M., et al. (2025). Exploring Bias in over 100 Text-to-Image Generative Models. arXiv. https://arxiv.org/html/2503.08012v1*
- *GSD Venture Studios. (n.d.). Cross-Cultural Considerations in AI Empathy. https://www.gsdvs.com/post/cross-cultural-considerations-in-ai-empathy*
- *Jain, P., et al. (2024). Prompting with Phonemes: Enhancing LLM Multilinguality for non-Latin Script Languages. arXiv. https://arxiv.org/html/2411.02398v1*
- *Lingoda. (n.d.). Lost in Translation: The Hidden Flaws of AI in Bridging Cultures. https://www.lingoda.com/blog/en/can-ai-truly-understand-culture/*
- *O'Brien, M., & Parvini, S. (2025, March 28). CHATGPT's Viral Studio Ghibli-style images highlight AI copyright concerns. AP News. https://apnews.com/article/studio-ghibli-chatgpt-images-hayao-miyazaki-openai-0f4cb487ec3042dd5b43ad47879b91f4*
- *OER Africa. (n.d.). Artificial Intelligence and the Underrepresentation of African Languages. https://www.oerafrica.org/content/artificial-intelligence-and-underrepresentation-african-languages*

- ***Pal, A., et al. (2024). Tokenization and Morphology in Multilingual Language Models: A Comparative Analysis of mT5 and ByT5. arXiv. https://arxiv.org/html/2410.11627v2***
- ***Penn State University. (n.d.). Non-Western cultures misrepresented, harmed by generative AI, researchers say. https://ist.psu.edu/about/news/venkit-aies-social-harms***
- ***Riou, M., et al. (2023). Linguistic disparities in cross-language automatic speech recognition transfer from Arabic to Tashlhiyt. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC10764819/***
- ***Siliezar, J. (2020, September 16). Joseph Henrich explores weird societies. Harvard Gazette. https://news.harvard.edu/gazette/story/2020/09/joseph-henrich-explores-weird-societies/***
- ***SmartOne.ai. (n.d.). The Must Read Guide to Training Data in Natural Language Processing. https://smartone.ai/blog/the-must-read-guide-to-training-data-in-natural-language-processing***
- ***Stanton, B., et al. (2024). AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. arXiv. https://arxiv.org/html/2409.11360v2***
- ***UNESCO Recommendation on the Ethics of Artificial Intelligence (2021). https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence***

Suggested Reading

- ***AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances: Research paper detailing the***

homogenizing effect of AI writing tools. https://arxiv.org/html/2409.11360v2

- *Investigating Cultural Alignment of Large Language Models: Explores how LLMs align with different cultural values. https://aclanthology.org/2024.acl-long.671.pdf*
- *Racial disparities in automated speech recognition: Study highlighting significant performance differences in speech recognition based on race. https://www.pnas.org/doi/10.1073/pnas.1915768117*
- *Non-Western cultures misrepresented, harmed by generative AI, researchers say: Discusses the misrepresentation and potential harm caused by generative AI. https://ist.psu.edu/about/news/venkit-aies-social-harms*
- *Artificial Intelligence and the Underrepresentation of African Languages: Focuses on the challenges and impact of underrepresentation of African languages in AI. https://www.oerafrica.org/content/artificial-intelligence-and-underrepresentation-african-languages*
- *AI Regulations: Global Trends, Challenges and Approaches: Overview of how different regions are approaching AI regulation, touching on cultural aspects. https://bowergroupasia.com/ai-regulations-global-trends-challenges-and-approaches/*