This is a forum for perspectives on designing for marginalized communities worldwide. Articles will discuss design methods, theoretical/conceptual contributions, and participatory interventions with underserved communities.
— Nithya Sambasivan, Editor

Seeing Like a Dataset from the Global South

Nithya Sambasivan, Google

Data is the fundamental technical infrastructure for inferential technologies. The Global South contributes an outsize user and labor base in producing the data that powers AI models. Yet most AI creators have failed to understand how the social, political, ecological, and infrastructural nuances of these contexts can affect data, including data quality, fairness and fair work, robustness, and model safety. Indeed, the centrality of data in building ML models is undergirded by assumptions of data objectivity, accuracy, and reliability in representing people and phenomena. In this article, I bring to light the unstated assumptions behind datasets that power AI models and examine them with alternative realities from the Global South, based on a series of recent research studies.

Data is available. Today, AI and ML technologies are relatively accessible to entrants from the Global South due to open-source and pretrained models, easy-to-access courses, and thriving practitioner communities worldwide. Disparities, however, show in data and computing resources [1]. With limited digital infrastructures and fewer socioeconomic datasets in the Global South, ready-made datasets are unavailable and data often needs to be collected from scratch. But AI education does not adequately prepare practitioners for real-world data work, instead focusing on toy datasets with clean values (e.g., UCI Census and Kaggle datasets). AI in practice requires the creation of data pipelines, often from scratch, going from ground truth to model maintenance.

State and industry apparatus in the Global South often collect and retain valuable, large-scale data, but the datasets are not always made publicly available due to infrastructure and nontransparency issues. For example, datasets featuring migration, incarceration, employment, or education data, divided by subgroups—so valuable to ML research areas like algorithmic fairness—are unavailable to the public. AI solutions have end-to-end opacity in the Global South, with unknown data, model behavior, and inferences. The AI imaginary can be aspirational, often rooted in hype and promise. AI-based solutions are readily adopted in high-stakes domains, often too early [2].

Data is representative. Datasets are often seen as reliable representations of populations, with biases in models frequently attributed to biased datasets, presupposing the possibility of achieving fairness by "fixing" the data. However, social contracts, informal infrastructures, and population scale in India lead us to question the reliability of datasets. Half the Indian population lacks access to the Internet—the excluded half is primarily women, rural communities, and Adivasis (indigenous people). Caste, class, and gender inequities may prevent the ability to access and create online content. For example, many safety apps use data mapping to mark areas as unsafe, in order to calculate an area-wide safety score for use by law enforcement; but safety apps can be populated by middle-class users, who tend to mark Dalit, Muslim, and slum areas as unsafe, potentially leading to hyperpatrolling in these areas. Data can also be missing due to artful user practices to manipulate algorithms, motivated by privacy, abuse, reputation, and other concerns [3,4]. Finally, "off data" practices (e.g., on the phone) can go undetected by conventional data-logging mechanisms, rendering them absent from datasets. Household dynamics can affect data collection, especially when using the door-to-door method. For example, heads of households, typically men, often answer data-gathering surveys on behalf of women, but their responses are recorded as women's.

Datasets that offer better goodness-of-fit to models, with parameters like mobility, expenditure, and literacy, largely correspond to data-rich profiles. In other words, models in or for the Global South may be overfit to digitally rich users, typically middle-class men. A model that is fair and equitable to various subgroups in the U.S. may cause damage to communities in the Global South, not simply because of the diversity of subgroups but also due to the correlation of "good data" with privileged demographics. For example, even

Insights

- → ML models need standardized, population-level datasets for operation, but both communities and infrastructures are highly diverse around the world.
- → Taken-for-granted assumptions around data availability, reliability, and representativeness all need careful examination against contextual realities.

IMAGE BY JMCM / SHUTTERSTOCK.COM

though decades of rigorous development economics research demonstrate that women are the most reliable loan borrowers and have shown exceedingly high repayment rates, they are highly marginalized when it comes to Internet access: Only 33 percent of Internet users in India are women. Despite their actual ability to repay loans, women get marginalized in loan-allocation AI systems that allocate better credit scores to men, due to properties like physical mobility. In turn, some women apply for loans using the accounts of their male relatives. The dataset definition and collection processes themselves can be limited and inaccurate, leading to models that overfit for specific demographics.

Another aspect is that marginalized communities in the Global South may have little to no recourse to AI data and models. The high-tech illegibility of AI can render accountability, contestability, and recourse out of reach for groups marginalized by literacy, legal, and educational capital. Even when feedback mechanisms are included in apps, prior work shows that they may be culturally insensitive or dehumanizing [3]. Human infrastructures like street-level bureaucrats, administrative offices, and frontline workers, who play a crucial role in providing recourse to marginalized Indian communities, are removed in AI systems.

Data is valued. Paradoxical to data's primacy, it remains the most undervalued and deglamorized aspect of AI system building. In our research on AI data practices employed by practitioners and researchers building high-stakes AI systems in parts of India, sub-Saharan Africa, and the U.S., we observed the sobering prevalence and severity of data cascades—compounding events causing negative, downstream effects from data issues [1]. Although the AI/ML practitioners in our study were attuned to the importance of data quality and displayed a deep moral commitment to vulnerable groups, data cascades were prevalent even in the high-stakes domains we studied, such as cancer detection and regenerative farming. Ninety-two percent of AI practitioners we interviewed experienced messy, protracted, and opaque data cascades. These cascades often resulted from the application of conventional AI practices that undervalue data quality. For example, eye disease–detection models, trained on noise-free data for high model performance, failed to predict the disease in production due to small specks of dust on images. Data cascades compounded into major negative impacts downstream of the models, such as costly iterations, discarding projects, and harm to communities.

Paradoxical to data's primacy, it remains the most undervalued and deglamorized aspect of AI system building.

The prevalence of data cascades point to the contours of a larger problem of broken data practices, methodologies, and incentives in the field of AI: residual conventions and perceptions in AI/ML drawn from worlds of "big data"—of abundant, expendable digital resources and worlds in which one user has one account; of model valorization; of moving fast to proof of concept; and of viewing data as grunt work in ML workflows. Additionally, our results point to serious gaps in what AI practitioners were trained and equipped to handle. These gaps come in the form of tensions in working with field partners and application-domain experts, and in understanding human impacts of models—a serious problem as AI developers seek to deploy in domains where governments, civil society, and policymakers have historically struggled to respond. For example, field partners, especially frontline workers who collected data, reported facing limited data literacy, poor pay, and information symmetry issues. In contrast, ML developers and top-level management often entered mutually synergistic partnerships through joint press releases or publications, leaving data workers marginalized.

Data cascades also reflect the effect of larger AI/ML field reward systems: Despite the primacy of data, novel model development is the most glamorized and celebrated work in AI—reified by the prestige of publishing new models at AI conferences, entry into AI/ML jobs and residency programs, and the pressure for startups to double as research divisions. Some practitioners reported feeling pressured to hack models for accuracy and performance, rather than being able to give due attention to quality, safety, or process.

Data is technical. As mentioned above, a rich human infrastructure [5] from public service delivery (e.g., frontline health workers) extends into AI data collection in the Global South.

Particularly in high-stakes domains and AI-for-social-good projects, datasets are not readily available; frontline workers often perform the labor of collecting datasets from scratch. Frontline workers take pride in and gain prestige from doing good work, which may include providing care to pregnant women or attending to infected crops. However, AI data collection can be orthogonal to their goals, workflows, or values and may come into conflict with their primary purpose: assisting communities.

In the Global South, data collection and curation often comprise human-mediated relations, in contrast to the popular conception of "automatic" or "technical" data collection. Human infrastructures point to questions around improving data literacy and provenance, as well as transparency on use cases, fairer incentives and work policies, collective and transitive consent, and social audits. To a data worker, for example, consent stems from interpersonal trust and holding others in high regard—relationships that may not be transmitted to downstream AI applications. In some cases, though, data workers have been shown to fudge data without having conversations with those affected; efforts like social audits and public hearings by the Mazdoor Kisan Shakti Sangathan have introduced better transparency and accountability.

Data categories are universal. Categories in datasets are assumed to remain static and uniform across the world, but they can actually carry very different nuances in meaning or implementation. Take the instance of proxies, which are used as substitutes for properties of protected groups. Proxies in India may be similar to those in the West but have entirely different implementation specifics, due to India being such a pluralistic country. For example, members of the Hijra community (a marginalized intersex or transgender community) may live together in a housing unit and be seen as fraudulent or invisible to models using family units. Proxies such as those for asset ownership may not generalize even within a country.

A name is the most semantically meaningful proxy in India, communicating caste, gender, religion, class, or ethnicity. Zip codes are heterogeneous, with housing of multiple socioeconomic classes abutting one another, in contrast to some homogeneous Western neighborhoods influenced by redlining in the past. Mobility has been reported to be much lower for women, due to personal safety concerns, and for people with disabilities, due to limited infrastructure such as ramps. Traditional occupations may correspond to caste or religion. AI systems in India remain underanalyzed for biases, mirroring the limited public discourse on oppression in this area, in contrast to, say, the anti-racism discourse.

Data is uniquely identifiable. AI system developers make a critical assumption so pervasive that it usually goes unstated: that user data corresponds one-to-one with unique individuals in the real world. This assumption gets challenged in the Global South through various sociocultural arrangements, including shared device and SIM card use, and the frequency with which people change their numbers. Location may not be permanent or tethered to a home; for example, migrant workers regularly travel across porous nation-state boundaries.

A way forward. Here are some guidelines to create more inclusive, representative, and ethically sourced data when working in the Global South:

- Does the data need to be collected? If yes, why? Does the affected community consent to the data collection, and what are their expectations from giving the data? Are we in a position to fulfill those expectations? Was anyone excluded in the dataset, and if so, what are the implications? Are the categories inclusive and supported by the affected communities?
- How did we define the parameters of our datasets? Have we combined observational research with domain expertise from communities, data collectors, and topical experts, to gauge whether the data is accurate and reliable? Do we have clear metrics for goodness of data, not simply goodness of fit?
- Did we work closely with the data collectors and partners in all stages of model development, and especially in defining the datasets? Did we provide the data collectors and annotators with transparency on downstream use cases? Did we design the processes and tools in collaboration with them and include an understanding of their workflows? Are we accountable to them?
- Do we have clear dataset documentation with specificities and assumptions outlined? Do we have a clear understanding of the contextual specificities where our models will be launched and whether our dataset is accurate, fair, or reliable?
- How can we equip ourselves to better understand the impact of our models on communities? Have we built clear ways for communities to contest and seek redress, assuming they were involved in the creation process?

It is important to note that these are general considerations for anyone working with data, anywhere. We should move from more data to *better* datasets and data work conditions.

Endnotes

1. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. 'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI. *Proc. of the 2021 CHI Conference on Human Factors in Computing Systems*.
2. Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. Re-imagining algorithmic fairness in India and beyond. *Proc. of the 2021 Conference on Fairness, Accountability, and Transparency*.
3. Sambasivan, N. et al. 'They don't leave us alone anywhere we go': Gender and digital abuse in South Asia. *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2019, 1–14.
4. Sambasivan, N. et al. 'Privacy is not for me, it's for those rich women': Performative privacy practices on mobile phones by women in South Asia. *Proc of the 14th Symposium on Usable Privacy and Security*. 2018, 127–142.
5. Sambasivan, N. and Smyth, T. The human infrastructure of ICTD. *Proc. of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. 2010, 1–9.

Nithya Sambasivan is a researcher at Google Research India, where she leads the HCI group. Her current research focuses on developing responsible AI by centering marginalized communities in the Global South.
→ nithyasamba@google.com

DOI: 10.1145/3466160