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

All Equation, No Human: The Myopia of AI Models

Nithya Sambasivan, Google

AI research and practice today places novel mathematical algorithms—models—at center stage, celebrating new model architectures and state-of-the-art performance. Although algorithms are trained over datasets and input by human experts, the overt emphasis on model performance ignores these aspects. The model emphasis has led to criticism from within the community, with some calling it "alchemy," "empirical challenges to be won," "incremental," and "leaderboards" [1]. While the AI/ML communities glorify models, dataset work has only recently been accepted at leading conferences, illustrating what has been considered science and what has not. In some cases, in order to report high performance in academic papers and funding venues, AI models are measured against large, clean datasets without noise, which is not representative of their performance in the real world [2].

The overt model emphasis is particularly problematic in the growing AI deployments in high-stakes domains with critical safety impacts on living beings. Several of these high-stakes AI projects seek to intervene in low-resource contexts, in fragile and complex domains—for example, cancer detection in rural Ghana. Several high-stakes AI projects are "high modernist," demonstrating strong confidence in the potential for scientific and technological progress as a means to reorder the natural world. For example, a sample vision statement from an AI researcher in our study was to diagnose tuberculosis from X-rays in 30 seconds, instead of 10 days, in low-resource regions. While the goal itself is noteworthy, these vision statements tend to measure technological efficiency while ignoring other metrics, such as the displacements and harms to stakeholders while achieving this goal.

I situate my research against this backdrop of North Star visions of AI in low-resource areas. Many people are employed in service of AI in low-resource areas, in the form of domain experts, annotators, and field partners. The downstream impacts of these AI projects on these communities can be huge. While my arguments are centered on AI development for low-resource populations, several of my points may ring true for any AI development that aims to intervene in risky and fragile areas.

Insights

- → Novel model development is celebrated in AI, at the cost of ignoring that algorithms are trained over datasets and input by human experts.
- → Economic analyses of AI are based on metrics that fail to measure how domain experts and workers were drafted into labor, and any displacements and redistributions caused by the model.
- → AI systems should aim for better transparency of surplus human labor, inclusion of data work, recognition of domain expertise in model building, and contextually appropriate safeguards.

TOO MUCH MODEL, TOO LITTLE BENEFIT

My colleagues and I empirically find [3] that the enormous focus on model development among developers and researchers stems from various incentives, such as publication prestige, how residencies are evaluated, competitive differentiation, and so on. AI education is now more accessible through courses and nanodegrees, but remains largely focused on model building, not addressing the real-world challenges of collecting data or deploying and measuring systems, which practitioners increasingly must do. The overt focus on models often comes at the cost of ignoring fundamental concerns around disempowered communities drafted into building or using these systems. As AI models increasingly seek to intervene in domains where governments, civil society, and policymakers have historically struggled to respond, this stance becomes problematic.

This view comes from the conventional AI/ML pipeline, which begins with an available, perhaps unclean, dataset and ends with model evaluation or deployment in a sanitized system. It is clear that the scope of what is considered AI must evolve. Metrics are often driven by concerns of "machine beats human" efficiency, cost, and outperforming industry standards, which often rely on the expertise of AI developers, rather than appropriate metrics, to evaluate claims relevant to the affected communities and phenomena, as decided by experts and communities. Until recently, non-model aspects like data, domain expertise, and meaningful safeguards for users have been considered to be outside of the scope of AI, relegated to "operations." The elite status of AI system building was restricted to developers, leaders of partner organizations, celebrity scientists, bureaucrats, and the machine intelligence itself.

Take data—it determines the performance, fairness, robustness, safety, and scalability of AI systems. Paradoxically, for AI researchers and developers, data work is often the least incentivized aspect, viewed as "operational" relative to the glamorous work of building models [3]. *Data cascades* are compounding events causing negative, downstream effects from data issues that result in technical debt over time. For example, eye-disease-detection models, trained on noise-free datasets to improve model performance, can result in total failure in predicting retinal disease when there are even small specks of dust on a camera lens. In our study, 92 percent of AI developers reported experiencing at least one cascade, with 45 percent experiencing two or more. Data cascades are opaque in diagnosis and manifestation, with no clear indicators, tools, and metrics to detect and measure their effects. They are triggered when conventional AI practices, such as viewing data as operations, moving fast, hacking model performance without consideration for data quality, and undervaluing domain expertise and labor, are applied in high-stakes domains. Data cascades have negative impacts on the AI development process, including harm to beneficiary communities, burnout of relationships, and the need to perform costly iterations.

Another important aspect of building AI systems—domain expertise—is often neglected, eliminated, and automated out in building these models [4]. Domain experts, such as community health workers and agricultural extension workers, are necessary for AI projects in areas with poor infrastructure where there are limited datasets available. These underpaid and overworked domain experts are often drafted to perform AI data collection for free, on top of their primary responsibilities. Despite the domain experts' mastery and knowledge that takes decades to build—and which the AI model seeks to emulate—we found that developers often reduced domain experts to mere data collectors for their expert models. Most developers did not have any firsthand contact with domain experts, much less provide them with training or compensate them for their data labor. Instead, developers attributed poor data quality to the poor work practices of domain experts, perceiving them as corrupt, lazy, noncompliant—and as datasets themselves. Domain experts were perceived as getting in the way of model-development efforts. To influence domain experts to collect better-quality data, developers created interventions built on these refractory associations, in the form of surveillance, gamification, cross-verification, and preprocessing fixes.

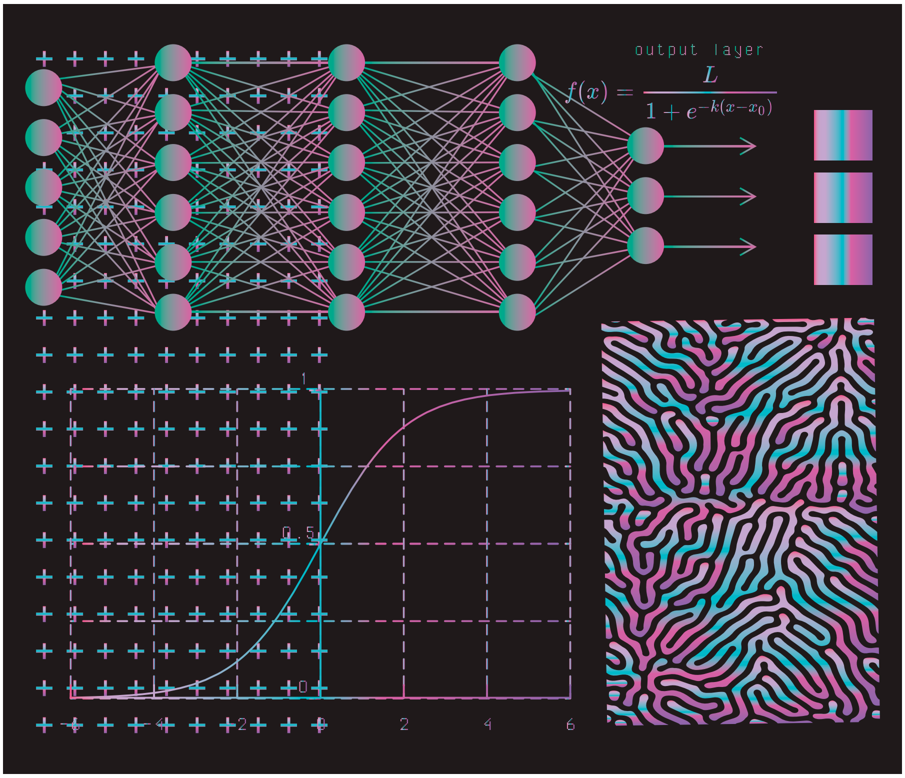

IMAGE BY LOCAL DOCTOR / SHUTTERSTOCK.COM

In this way, AI development risks fundamentally deskilling domain experts in low-resource contexts. Even though the models in our study sought to emulate and improve over the expertise of domain experts, the experts' knowledge was treated as nonessential for models. AI developers, who are experts in their technical fields but not in the application domains—for example, a developer building a cancer-prediction model—will inevitably leave gaps when domain experts are excluded from the model that seeks to learn their knowledge (through datasets).

A final crucial aspect of building AI systems, algorithmic fairness, has remained model-centric until recently [5]. While optimizations like predictive parity for different subgroups are important, they are often reductive fixes, and especially do not always scale to non-Western contexts, in which they did not originate. Conventional algorithmic fairness makes several assumptions, including: that fixing the data leads to fixing the bias, that users can benefit from the model fixes, and that there is a surrounding environment of accountability and regulation to address fairness. Several of these assumptions break in non-Western contexts such as India and Mexico, where responsible AI policies from the West are often copied and pasted. We argue that the distance between models and the disempowered communities in the Global South whom we hope to serve is large, due to factors like literacy, legal capital, and income inequality. A myopic focus on localizing "fair" model outputs alone can backfire. First, data can be missing in the form of digital divides or demographic inequities. Proxies may not generalize within pluralistic and diverse populations. Second, countries in the Global South are sometimes seen as sandboxes for industry and academia alike, deploying low-quality products with intrusive data collection and no recourse for these communities. Finally, some nation-states view AI aspirationally, leading to the introduction of high-stakes deployments, often with little testing or regulation. As a whole, the emphasis on models in algorithmic fairness offers a veneer of credibility to system builders, but when examined closely, these frameworks can be dangerously symbolic, non-generalizable to non-Western contexts.

The following are ways to better include humans in real-world AI systems:

Expertise and surplus labor. How expertise is defined, who is considered an expert, and how the expertise fed into models is recognized and credited are all important questions to answer. We impress upon AI developers the need to embrace more participatory stances that involve humans. Unregulated surplus labor—those involved in data collection or system deployment (including domain experts)—poses new questions about who the contributors to AI are and how they should be recognized. There are long chains of humans who are involved in enabling model development in AI in low-resource areas. How should coauthorship, attribution, transparency, provenance, and compensation evolve to include the surplus labor?

Data work. Our results indicate the sobering prevalence of messy, protracted, and opaque data cascades in high-stakes domains. Data cascades point to the contours of a larger problem: residual conventions and perceptions in AI/ML drawn from the worlds of big data—of abundant digital resources, of model valorization, of moving fast to proof of concept, and of viewing data as grunt work in ML workflows. We need to move toward a proactive focus on the practices, politics, and values of workers in the data pipeline. We need to move from the current approach of goodness of fit to goodness of data, from doing more to doing better with data.

We need to innovate further on structural incentives to recognize data work, in conference tracks (NeurIPS's Datasets track is a welcome start; https://neurips.cc/Conferences/2021/CallForDatasetsBenchmarks), organizational recognition of data work, greater collaboration and transparency with data collectors and domain experts, and more. Data ethics and practical data work, oversight boards like IRBs, and ethics standards should be a part of AI education and practice.

Partnership. Even though the models in our study sought to emulate and improve over the expertise of domain experts, the experts were limited to instrumental data collection and treated as nonessential knowledge for models. AI developers, who are experts in their technical fields but not in the application domains, will inevitably leave gaps when domain experts are excluded from the model that seeks to learn their knowledge. The data-quality issue is only an issue if we think of domain experts in these limited ways. But if we were to reimagine domain expertise as an essential partnership throughout the AI pipeline, we could see new possibilities for collecting, modeling, and scaling knowledge. Domain experts can contribute to critical questions that can affect model behaviors: What exactly are we modeling? What assumptions are appropriate? What features should be included in the model? What are we trying to predict? How will we know? Instead of motivating an overworked health worker to do more work for dataset collection, one might ask how to help them achieve their goals better, such as by prioritizing their numerous visits, better capturing their in situ knowledge, better allocating limited medical resources, and building visibility into their contributions. We need better recognition and attribution of domain experts' contributions.

Responsibility. A responsible AI strategy for low-resource areas needs to reflect the deeply plural, complex, and contradictory nature of these contexts and needs to go beyond model fairness. Due to the data and model distortions, we must combine datasets with an understanding of context. The thriving human infrastructures point to new ways of looking at data as dialogue. Marginalized communities need to be empowered in identifying problems, specifying fairness expectations, and designing systems to avoid top-down fairness. Contextual heterogeneity means that fair ML researchers' commitment should go beyond model outputs to creating systems accessible to those communities. Unequal standards, inadequate safeguards, and dubious applications of AI in low-resource areas may lead to harmful effects.

Economic analyses. Economic measurements of AI, such as the vision statement on detecting TB in 30 seconds instead of 10 days mentioned earlier, measure the before-and-after effects of algorithms dropped into a social setting. However, they miss out on measuring how domain experts and workers were involved and drafted into hidden labor, how experts and communities were displaced as a result, and how expertise was redistributed as a result of the model. We need to expand the parameters of what gets measured and seen to fully understand the effects of these AI systems.

Endnotes

1. Heinzerling, B. NLP's Clever Hans moment has arrived. *The Gradient*. Aug. 26, 2019; https://thegradient.pub/nlps-clever-hans-moment-has-arrived/
2. Marcus, G. Deep learning: A critical appraisal. arXiv preprint, 2018; arXiv:1801.00631
3. 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 CHI 2021*. ACM, New York, 2021.
4. Sambasivan, N. and Veeraraghavan, R. Deskilling of domain expertise in AI. Under review.
5. Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. Re-imagining algorithmic fairness in India and beyond. *Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, 2021, 315–328; https://doi.org/10.1145/3442188.3445896

Nithya Sambasivan is a research scientist at Google Research and leads the HCI group at the India lab. Her longstanding research on marginalized communities in the Global South has deeply shaped Google products and received numerous awards at top conferences.
→ nithyasamba@google.com

DOI: 10.1145/3516515