

AI's not fAIr

With AI predictions increasingly impacting people's lives, developers are looking at biases in models, and working to eliminate stereotypes.

SWETA AKUNDI

It became increasingly clear to venture capitalist Raji Baskaran as she surfed the net that Google preferred 'Hari' to 'Hamsa'. Baskaran had done a quick search to show Hamsa Balakrishnan's work profile to an investor interested in a start-up run by the Massachusetts Institute of Technology (MIT) Professor. But instead of Hamsa, the search engine invariably threw up the name of her brother, Hari, also a Professor at MIT.

This experience in the summer of 2019 in Portland, U.S., drew Baskaran's attention to a yawning gender gap in data scraped from the internet. So, in 2022, she set up Superbloom Studios to bridge these gaps in the digital world. The company is now

working on a long-term, open-source project called Hidden Voices, which seeks to reduce gender biases in search algorithms by adding 10,000 women's biography drafts to a not-for-profit library such as Wikipedia.

Over 50% of Wikipedia users are women, but only 15% of its editors are women – and fewer than 20% biographies are of living women, points out Baskaran. Search algorithms develop a bias based on statistics to give “relevant” results, she adds. Better-represented groups (men in science, in this case) are likely to show up more in searches, and, consequently, likely to be quoted more, and thus increase their representation further. “It's a chicken-and-egg problem,” Baskaran says.

With Hidden Voices, she hopes to ensure that the available data for artificial intelligence (AI) models is better represented. “With the help of expert volunteers who point us to credible sources, we use ML (machine learning) models that can collate data to create Wikipedia-style biographies of underrepresented minorities, starting with women in STEM (Science, Technology, Engineering, and Mathematics),” she says.

The impact of human biases seeping into AI extends beyond search algorithms. Predictive ML models today greatly help in making sound judgements. Even before humans enter the decision-making loop, AI foretells a person's chances of, say, being shortlisted for a job, being eligible for a housing loan, or even developing a certain type of cancer.

With AI outcomes influencing decisions, ‘fairness’ is a key metric to be added to a model's parameters, along with accuracy, size and efficiency. To ascertain whether or not a model is fair, a person would need to understand how AI algorithms arrive at their outcomes.

With this as part of its mission, the Centre for Responsible AI (CeRAI) was set up in Chennai this year. “Machine learning

works because it is able to generalise by making some assumptions about what is similar and what is dissimilar,” says B. Ravindran, head of CeRAI. Biasing, in ML, is classifying: a way to learn about the world, through experience.

“The fact that two people of the same age group, who work in the same place, and like the same author, might like the same new book, is not offensive. But when you say that two people of the same ethnicity will only like a certain kind of food, that becomes problematic,” Ravindran explains. And in India, with its many ethnicities and religions, even the idea of what is problematic might change from State to State, district to district. “With the diversity in India, a lot of the bias is not codified — it is implicit. You know a stereotype when you see it, but there isn't a legal characterisation of what bias is.”

Representing diversity

Ravindran's team at CeRAI is looking at building a ‘stereotype’ dictionary relevant to the Indian context. This is similar to the global ‘StereoSet’ built in 2020 by developers from MIT, Intel, and the Canadian AI initiative CIFAR to evaluate bias on the axes of gender, race, religion, ability and profession. “It is a stereotype dictionary that you can plug into a model so that it optimises not learning those stereotypes; or, if it learns it, it gives it less weightage,” Ravindran says.

Search engine company Google has taken a special interest in removing the West-centric lens to algorithmic fairness by funding studies such as ‘Re-imagining Algorithmic Fairness in India and Beyond’ (bit.ly/Fairness-India). When CeRAI was launched, it received funding of \$1 million

CeRAI and Google India are working on Project Bindi, a framework that evaluates and mitigates fairness issues in publicly available natural language processing models.



from Google to provide an Indian context to AI fairness.

CeRAI and Google India are working on Project Bindi, a framework that evaluates and mitigates fairness issues in publicly available natural language processing (NLP) models. Leading this effort is Google developer Shachi Dave, who, in 2022, published a paper on ‘Re-contextualizing Fairness in NLP: The Case of India’ (bit.ly/fairness-NLP). In this paper, Dave and co-authors first agree on the prominent axes of social disparities in India: gender, region, religion, and caste — and then outline the positive and negative associations with each group along these axes. For this, data was collected online as well as offline.

“We complemented the LLM (large language model)-based approach with a community approach, through an outreach to eight suburban and urban colleges to understand the stereotypes people experience in daily life,” Dave detailed at a presentation of her work at the Indian Institute of Technology (IIT) Madras in April 2023. The community outreach helped them identify local stereotypes and slang that would not be understood by LLMs online. In these interviews, they found that college-goers often perceived a particular

community as brave, another as artistic, and so on.

Bias in LLMs shows up in image-generative software as well. In August 2023, a team of researchers from University of California Santa Cruz created the ‘Text to Image Association Test’ under the guidance of Xin Eric Wang, Assistant Professor of Computer Science and Engineering, to quantitatively measure biases embedded in text-to-image models such as Stable Diffusion.

When you ask a generative AI model to show you a picture of a ‘person acting as a caregiver’ or ‘a person working in a science laboratory’, what are the likely gender and race of the subject in the results? Would the results, based on the wording of the prompt, be skewed to a particular gender and race? These are the questions that the tool can answer.

It relied on the image-text paired datasets scraped from the web. The researchers explain that bias occurs in this selection when data is not suitably collected from a diverse set of data sources, or the sources themselves do not adequately represent different groups of populations.

Illustratively, it has been reported (bit.ly/India-context) that nearly half of the samples of ImageNet, which is the base for most deep learning of visual data, come from the United States, while China and India, the two most populous countries in the world, contribute only a small portion of the images.

Toolkits to audit fairness

Making the data inclusive is the first step towards fairer AI. Or, suggests Chennai-based researcher Gokul Krishnan, depending on the use case, you could improve data anonymisation — that is, remove all identifiers and tags from the source data. Krishnan, who recently spent six months studying bias in healthcare NLPs at the U.S. National Institute of Standards and Technology, says that in one particular case, it was discovered that a model was more likely to diagnose symptoms correctly if the patients had a certain health



Even before humans enter the decision-making loop, AI foretells a person's chances of, say, being shortlisted for a job or being eligible for a loan.



CASE IN POINT

All are equal in the eyes of law. Does Legal AI agree?

Since 2021, the Supreme Court of India has been using an AI-controlled tool designed to process information and make it available to judges for decisions. Although it has no role in the decision-making process, AI can help generate and translate case summaries and retrieve cases with precedence from reams of archives (see *'Courting AI, legally!'*, bit.ly/shastra-legal).

A projects undertaken by the Chennai-based Centre for Responsible AI (CeRAI) tested fairness in InLegalBERT, a language model that trained legal AI such as LegalBERT on Indian legal text. "We tested for bail prediction, and found that the model predicted one religion to be more likely to get bail on murder charges, whereas a certain other community is more likely to get bail on a dowry charge. Changing just the religion changed the prediction," says B. Ravindran, head of CeRAI. "This is not to say that the legal system is biased," he emphasises, "but that if deployed without testing, legal models can learn false biases based on skewed data."

When compared to vanilla BERT, InLegalBERT showed a greater gender bias: a doctor was more closely associated with being a man than a woman; a nurse with a woman. "The higher gender bias could be because a lot of legalese in India is not written in a gender-neutral manner," Ravindran reckons.

insurance provider. "That was because the data had not been anonymised enough. The model was trained on cases that mostly had this particular insurance, which the developers had not considered as an attribute to be anonymised," Krishnan elaborates.

However, it is difficult to ascertain how much anonymity is enough. Models pick up on other correlations that enforce the same bias. You could anonymise race, but as some communities are more likely to live in certain areas, the bias would shift to PIN codes instead.

With all these evolving metrics in mind, companies such as IBM, Microsoft and Fiddler AI have developed toolkits that audit models for fairness. These are deployed to maintain internal standards as well as sold to other business enterprises using AI.

Bengaluru-based engineer and Senior Manager at IBM Research AI, Sameep Mehta, is a part of the team conceptualising the company's AI Fairness 360 (AIF360) toolkit. "When we started the work in 2017, we realised that we can't build trust and transparency tools behind closed doors; it has to be a community effort," Mehta says.

AIF360, a Python-language tool, was made available on GitHub for developers to collaborate on and provide open-source algorithms for bias detection and mitigation. The toolkit measures fairness on 70 metrics using ten types of algorithms.

For business enterprises, this means being able to test their model's fairness score before deploying it. "It is up to you to decide which metric to use to measure bias depending on whether you are providing inclusive data or want to account for sampling bias; we provide a guide for that. There are also point-and-click demos that train classifiers on the go to explain how biases arise," Mehta says.

The choice of retaining certain biases lies with the enterprise. "We don't dictate what is a stereotype and what is not. What we are saying is that we are going to provide tools which will help you discover bias in your model," Mehta says. The enterprise has to decide which of these biases it wishes to fix, he says.

A bank using an AI model to decide people's credit score would, by default, need to discriminate between people. However, if

the model is rejecting a loan based on just one protected attribute — say the person's address or gender — the bank may want to fix it to prevent the loss of potential customers. At the same time, the bank might have a special offer for new customers, in which case the model would need to be biased toward them as a marketing strategy.

Microsoft has a similar open-source platform on GitHub called Fairlearn, co-founded by Senior Principal Researcher Miro Dudík. In an online webinar organised by Microsoft, Dudík uses the example of AI assistants' established ability to understand some communities better than others (based on dialects, age and so on) to talk about Fairlearn's 'reductions approach' of mitigating algorithmic unfairness. The metrics for fairness here would include measuring the error rate for different groups and calculating the largest difference between the error rates of two groups (ideally close to 0), or the smallest ratio between two groups (ideally close to 1). The 'reductions approach' can be applied to pre-existing AI models post-processing, Dudík explains.

"We start off by training the model on initial data and then check if the fairness constraints are violated. If they are, the tool reweighs the data

in a specific fashion and sends it back to the standard algorithm, and it reiterates so on," Dudík says. "This means that you can take an existing algorithm which you have previously used to train your systems, combine it with the reduction approach to now obtain a model that optimises performance while also satisfying fairness constraints."

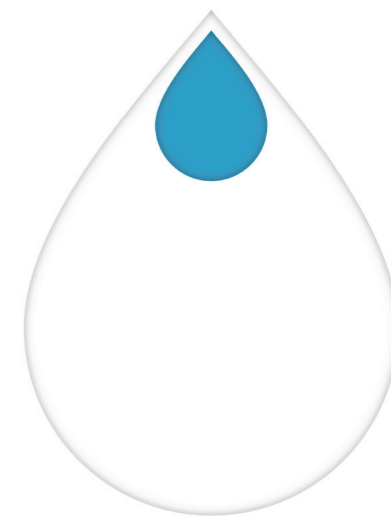
According to research at IBM, lack of trustworthiness in AI is the biggest impediment to its adoption in India, Mehta says. "If any enterprise or application deploys biased models only to be revealed as unfair later on, it puts AI adoption two steps back."

The road to trusting AI is indeed a long one. Today, four years after Baskaran first noticed Google's preference for Hari over Hamsa, she points out that the anomaly has been fixed. Hamsa now gets her own summary card, complete with links to her LinkedIn profile and to a YouTube video of one of her talks. Removing biases, clearly, is a work in progress — improving every day. •

Companies such as IBM, Microsoft and Fiddler AI have developed toolkits that audit models for fairness.



Rethink Consumption .

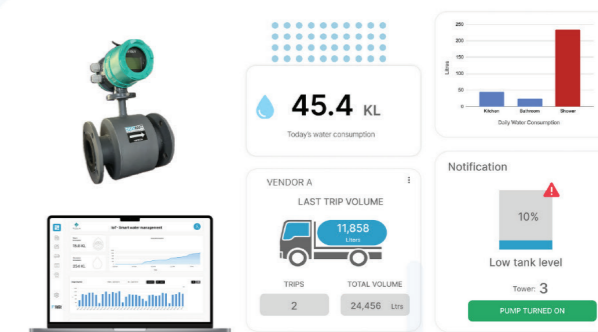


Water Saving Fixtures

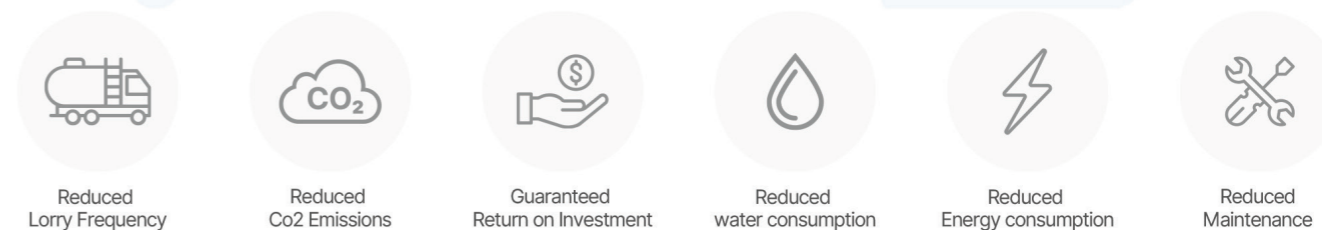
Achieve unparalleled water savings of up to 95% with our comprehensive range from retrofit solutions to specialized home kits. Sustainably machined from lead-free brass for safe, long-lasting durability.

Smart Water Management

EarthFokus Smart Water Management combines real-time data, automation, and scalability to revolutionize water usage, making it both cost-effective and eco-friendly.



True Impact



Save Beyond Water: With every drop saved, contribute to a broader narrative of holistic conservation and responsible living.