Research Statement

Avijit Ghosh
ghosh.a@northeastern.edu

My research lies at the confluence of machine learning (ML), ethics, and policy. In terms of research focus, I seek to discover and solve technical challenges in implementing Fair ML algorithms from theoretical research into the real world. I tackle this with a combination of methods: Deep Learning foundations, Data Mining, and Quantitative Measurement and Analysis, working in an interdisciplinary fashion with not just fellow computer scientists, but also with philosophers, lawyers, and investigative journalists. My long-term research goal is to clearly identify all the pain points in the deployment pipeline of pervasive algorithms, and suggest actionable solutions for both practitioners and policymakers, so that everyone in society can begin to benefit equitably from the massive strides in AI research.

So far in my research career, I have produced a total of 14 research papers, with at least 4 ongoing manuscripts. I have published at top ML and AI Ethics venues, such as FAccT, AIES, SIGIR, Neurips, and AAAI. Additionally, I have been a reviewer at many of these venues, and have organized workshops and panels as a member of QueerInAI. My work has also been covered by the press, notably in *Propublica*, *Wired*, *The Federal Reserve*, and the *MIT Tech Review*. I have been invited to speak at seminars, organized by places such as the *Trustworthy ML Initiative* and *the UK Government Centre for Data Ethics and Innovation*. I genuinely enjoy investigating algorithmic harms and solutions, and educating the broader community—both practitioners and regulators—about my discoveries.

Context

Algorithmic decision is permeating modern life, including high-stakes decisions like credit lending, bail granting, hiring, etc. While these ML models are great at scaling up processes with human bottlenecks, they also have the unintended problem of embedding unfair social biases like racism, sexism, homophobia, ableism, ageism, and religious intolerance.

In response to this, there is a growing body of academic work on ways to detect algorithmic bias and develop classes of fair algorithms. For example, there is now extensive literature presenting techniques for training fair classification and ranking models. Companies are beginning to adopt and deploy fair ML systems in real-world contexts, and lawmakers around the world have begun to take steps to regulate and enforce responsible ML in deployment, for example the Algorithm Accountability Act[1] and the Blueprint for an AI Bill of Rights[2] in the United States, and the AI regulatory framework[3] in Europe.

As with most burgeoning disciplines, however, fair ML research faces the challenge of developing methods and techniques that work *in situ*, but may fail during implementations in the real world. My work has identified a number of these gaps, including: western-centric fairness notions that may exhibit poor intersectional coverage [6], the lack of high quality demographic attributes at runtime that are required by many fair models [5], the lack of adversarial robustness of fair models [2], and the unfairness of fair models over time [7].

In my work, I specifically focus on these implementation-time challenges of fair ML models. My goal is to demonstrate the severity of these issues in practice, to heighten awareness among researchers, practitioners, and regulators, and present solutions to these real world implementation challenges.

Research Highlights

I have tackled several real-world challenges of deploying Fair ML research in my work.

Lack of Demographic Attributes during Fair ML Deployment

In cases where people are the data subjects being input to classification or ranking algorithms, the vast majority of existing work assumes that ground-truth demographic information will be available. Unfortunately, this assumption about the

[1] https://www.congress.gov/bill/117th-congress/senate-bill/3572
[2] https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[3] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

availability of ground-truth demographic data is often violated in practice. For example, in real-world contexts like assessing job applicants or credit seekers, social and legal barriers may prevent algorithm operators from collecting peoples' demographic information. The unavailability of ground-truth demographic data has led some system developers to adopt an alternative approach: infer protected class information from data and then supply it to the fair algorithm as input. One example of this is the Bayesian Improved Surname Geocoding (BISG) inference algorithm that is used by lenders and health insurers in the U.S. to infer people's race and ethnicity.

In my SIGIR 2021 paper **"When Fair Ranking Meets Uncertain Inference"** [5], I examine the following research question: *How does noise in demographic information as an input to a fair ML algorithm adversely impact the intended fairness of the outcomes for different subgroups?* Specifically, I test this for fair ranking. With a combination of simulation studies and 3 real-world ranking datasets, I establish the following two findings: (1) In the simulation study, the fairness metrics (both representation-based and exposure-based) of the final ranked list increased monotonically with the increase in the accuracy of the prediction of the protected attributes. However, the relevance of the ranked list, as measured by NDCG, barely changed, signifying that it is possible to perform significant fairness interventions without noticeably affecting the quality of rankings, and (2) I observed that the different rate of mispredictions for different demographic groups led to not only less fair rankings than if there were no noisy labels, but for certain demographic subgroups, the "fair" rankings were actually even more unfair than if no fairness intervention was performed. This is an alarming finding, showing that if not operationalized correctly, a fair algorithm can selectively perpetuate unfairness.

Adversarial Attacks can make Fair ML Unfair

Another serious concern in the ML community is model *robustness*, especially in the face of clever and dedicated adversaries. The field of adversarial ML has demonstrated that seemingly accurate models display surprising brittleness when presented with maliciously crafted inputs, and that these attacks impact models across a wide-variety of contexts. The existence of adversarial ML challenges the use of models in real-world deployments, particularly deep learning models. In my FAccT 2022 paper **"Subverting Fair Image Search with Generative Adversarial Perturbations"** [2], I explore the intersection of these two concerns—fairness and robustness—in the context of ranking: *when a ranking model has been carefully calibrated to achieve some definition of fairness, is it possible for an external adversary to make the ranking model behave unfairly without having access to the model or training data?* In other words, can attackers *intentionally* weaponize demographic markers in data to subvert fairness guarantees?

To investigate this question, I present a case study in which I develop and then attack a fairness-aware image search engine using images that have been maliciously modified with trained Generative Adversarial Perturbation (GAP) models. To strengthen my case study, I adopt a strict threat model under which the adversary cannot *poison* training data for the ranking model, and has no knowledge of the ranking model or fairness algorithm used by the victim search engine. Instead, the adversary can only add images into the query database, *after* the image retrieval model is trained. I observe the following findings from extensive experiments (1) The attacks can successfully confer significant unfair advantage to people from the majority class (light-skinned men, in my case study)—in terms of their overall representation and position in search results—relative to fairly-ranked baseline search results. (2) The attack is robust across a number of variables, including the length of search result lists, the fraction of images that the adversary is able to perturb, the fairness algorithm used by the search engine, the image embedding algorithm used by the search engine, the demographic inference algorithm used to train the GAP models, and the training objective of the GAP models. (3) The attacks are *stealthy*, i.e., they have close to zero impact on the relevance of search results.

A Fair ML Model can become unfair over time

Machine Learning (ML) and Artificial Intelligence (AI) models that are deployed into the field cannot guarantee consistent performance over time. One of the reasons for this might be that the underlying data has changed stochastically. This phenomenon, called *drift*, has been well-studied in the literature, from sudden to gradual drifts. Drifts may also be caused by true shifts in the relationship between the underlying variables (e.g., due to changes in the population over time), sampling issues, or even bugs that impact downstream data collection. In scenarios where a deployed ML model is making sensitive decisions, I argue that analyzing the impact of drift on the *fairness* of the model is equally, if not more, important than assessing the impact of drift on traditional performance metrics like accuracy and recall.

In my AIES 2022 paper **"FairCanary: Rapid Continuous Explainable Fairness"** [7], I present FairCanary, a continuous model monitoring system that offers two significant, novel capabilities versus state-of-the-art commercial systems that help ensure model fairness over time. (1) FairCanary incorporates a novel model bias quantification metric called Quantile

Demographic Disparity (QDD) that uses quantile binning to measure differences in the overall prediction distributions over subgroups. Because QDD is measured over continuous distributions, it does not require developers to choose specific (and often ad hoc) thresholds for measuring fairness, unlike most conventional fairness metrics. Additionally, QDD does not require outcome labels, which may not be available at runtime. (2) FairCanary reuses explanations computed for each individual prediction to quickly compute explanations for its bias metrics. This optimization makes FairCanary an order of magnitude faster than previous work that has tried to generate feature-level bias explanations.

Fairness Metrics and Measurements: Intersectionality, Inequality, User Choice

Most existing fairness metrics deal with either a binary view of fairness (protected vs. unprotected groups) or politically defined categories (race or gender). Such categorization misses the important nuance of intersectionality - biases can often be amplified in subgroups that combine membership from different categories, especially if such a subgroup is particularly underrepresented in historical platforms of opportunity. In my AAAI workshop paper **"Characterizing Intersectional Group Fairness with Worst-Case Comparisons"** [6], I present a method to convert binary fairness metrics into intersectional versions by a simple worst-case analysis, where compare the ratios of the metrics for the least advantaged group with the most advantaged group, with the aim of equalizing the gap so that every group is treated equitably.

My interest in fairness metrics was also of use during my two internships in the Twitter ML Ethics, Transparency and Accountability (META) team. During my first internship, I developed alternatives to conventional fairness metrics that required demographic information, that a company the scale of Twitter did not have access to for all users. To do this, I drew from distributional inequality measures in economic theory, and through extensive experiments, discovered that inequality metrics like GINI or Atkinson's Index appeared to be highly correlated with Demographic Disparity metrics. The second internship project was to make the home timeline more diverse without impacting core business metrics, my task was once again to come up with novel metrics that did not require user's private demographic information and yet made the home timeline more fair while also keeping in mind user's follow graphs. All the metrics I developed over my two internships were also shown to be orthogonal to Twitter's core business metrics, and therefore were ideal candidates for deployment in Twitter's internal responsible ML workbench.

Investigating Emerging Technology

Finally, my interest in how algorithms shape society in the real world can be seen from my different audits of real world systems. I have conducted several investigative research projects about Facebook's ad ecosystem, showing that wealthy advertisers on the platform use more privacy-sensitive targeting techniques [4], and how their Special Audiences algorithm, designed to remove demographic sensitive attributes from targeting lists, were still effectively biased [1]. The latter work was covered in the media [4] and ultimately led to Facebook completely revamping their targeting system and sunsetting the problematic algorithm[5].

I have also investigated Pymetrics's hiring algorithm for bias as part of a cooperative audit [8], and this informed legislation about Algorithmic Fairness in New York City[6]. While valid concerns remain that the regulation does not go far enough, it is a first step towards solutions-based research work helping shape policy. Currently, I am looking at whether Google/Bing Image search pushes people into filter bubbles via representational biases. I have started to investigate the emerging harms of Generative AI Art [3], especially how they shift profits from the hands of individual artists into the hands of corporations that own these models, as well as rampant plagiarism and lack of consent during model training.

Research Agenda

I would like to continue on my path of exposing shortfalls of real world technical systems, and designing solutions for fairer ML that works in both theory and practice.

Adversarial Defenses against fairness attacks: I have already shown that it is possible to attack fair ML models via adversarial attacks to cause it to become unfair [2]. I plan to investigate solutions to such an attack, through a mix of methods such as adversarial training, cloaking defenses, and online learning with bias annotations - to find a solution that is

[4] https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement

[5] https://about.fb.com/news/2022/06/expanding-our-work-on-ads-fairness/

[6] https://apnews.com/article/technology-business-race-and-ethnicity-racial-injustice-artificial-intelligence-2fe8d3ef7008d299d9d810f0c0f7905d

both computationally scalable and is robust against not just blackbox attacks, but hopefully also stronger whitebox attacks.

Machine Unlearning to remove problematic training data: I have taken a keen interest in the rampant phenomenon of massive models being trained on data scraped without consent – this includes both Large Language Models (LLMs) and also Text-to-Image Generation models. These models have been shown to not only be riddled with sexist, racist, toxic, or factually incorrect content in their outputs, that the model creators themselves concede are too difficult to correct for, but in the case of text generation models such as Github Copilot[7] and Image generation models such as DALL.E[8], it also presents the deeply concerning aspect of models trained on unconsenting individuals' data and then using them for profit [3]. Retraining such massive models from scratch is a prohibitingly expensive task, and I believe building on early research work in Machine Unlearning - to build a tool for individuals to request model owners to "forget" their training data and respect their IP, is the more feasible technical solution to this predicament. This is unexplored and certainly ambitious, but I plan to work with both Machine Learning scientists in academia and companies who are in the business of commercializing these models to find actionable, cheaply computatble and scalable unlearning solutions.

Injection of bias via Human Stakeholders: Humans can impact the behavior of a ML pipeline in two ways – bias added by annotators in the annotation stage before model training, and personal bias of the decision makers who are in charge of converting predictions into outcomes. While creating machine learning datasets, it would be interesting to measure annotator perspective and cultural bias in annotations - whether the gaps in perception are different for different demographic groups. Normatively speaking, I would like to think more deeply about whether a person's self disclosed attribute is the more important factor in corrective fairness techniques than what the majority of annotators think is their label. In terms of decision making bias, I would like to delve deeper into how humans in charge of making the final decision might subvert the algorithmic fairness interventions – for instance human recruiters on the other end of a fair candidate ranking system. This would involve working with people in human computer interaction, labor economics, and psychology.

Broader Impact

Helping Develop Actionable Policy: I would like my solution-focused research approach to aid better regulation of AI/ML practice. I plan to initiate dialogue with policymakers and regulatory agencies worldwide, such as the FTC and CFPB in the US or analogous bodies within the EU Commission, to help fine tune policy. My belief is that my work will be able to highlight specific technical interventions that model operators can take to implement AI responsibly, and such specific regulation will avoid cases of escaped accountability due to vague regulatory language. Ultimately, I would like my research to shape policy and have lasting positive change in society.

Publications

[1] Piotr Sapiezynski, **Avijit Ghosh**, Levi Kaplan, Aaron Rieke, and Alan Mislove. Algorithms that" don't see color" measuring biases in lookalike and special ad audiences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 609–616, 2022.

[2] **Avijit Ghosh**, Matthew Jagielski, and Christo Wilson. Subverting fair image search with generative adversarial perturbations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 637–650. Association for Computing Machinery, 2022. ISBN 9781450393522. doi: 10.1145/3531146.3533128.

[3] **Avijit Ghosh** and Genoveva Fossas. Can there be art without an artist? *arXiv preprint arXiv:2209.07667*, 2022.

[4] **Avijit Ghosh**, Giridhari Venkatadri, and Alan Mislove. Analyzing political advertisers' use of facebook's targeting features. In *IEEE workshop on technology and consumer protection (ConPro'19)*, 2019.

[5] **Avijit Ghosh**, Ritam Dutt, and Christo Wilson. *When Fair Ranking Meets Uncertain Inference*, page 1033–1043. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. URL https://doi.org/10.1145/3404835.3462850.

[6] **Avijit Ghosh**, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In Deepti Lamba and William H. Hsu, editors, *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, volume 142 of *Proceedings of Machine Learning Research*, pages 22–34. PMLR, 09 Feb 2021. URL https://proceedings.mlr.press/v142/ghosh21a.html.

[7] **Avijit Ghosh**, Aalok Shanbhag, and Christo Wilson. Faircanary: Rapid continuous explainable fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 307–316, 2022.

[8] Christo Wilson, **Avijit Ghosh**, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677, 2021.

[7] https://github.com/features/copilot
[8] https://openai.com/dall-e-2/