

# Avijit Ghosh

Applied Policy Researcher, ML & Society, Hugging Face

☎ (+1) 857-337-0180 | ✉ [avijit@huggingface.co](mailto:avijit@huggingface.co) (Work) | [avijitg22@gmail.com](mailto:avijitg22@gmail.com) (Personal) | 🌐 [evijit.io](https://evijit.io) | [in](#) [evijit](#) | [🔗](#) [evijit](#)

Algorithmic Fairness Ethical AI Machine Learning AI Explainability Policy Computational Social Science

## Education

### Northeastern University

*Ph.D. in Computer Science (Advised by Dr. Christo Wilson)*

Boston, MA

2019 - 2023

### Indian Institute of Technology (IIT) Kharagpur

*B.Tech. in Chemical Engineering, M.Tech in Financial Engineering, Minor in Computer Science*

Kharagpur, India

2014 - 2019

## Doctoral Thesis

### Algorithmic Fairness in the Real World: Challenges and Considerations

Defended June 2023

- Social bias in machine learning algorithms is a widespread problem that has been addressed through various measures, but implementing fair machine learning systems in the real world is challenging due to issues like noisy demographic information, adversarial vulnerabilities, policy restrictions and complex interactions between humans and algorithms.
- In my thesis, I attempt to outline these problems in fair ML systems, with the aim to gain a more complete understanding of the challenges involved and to be able to provide technical and policy recommendations to overcome their real world implementation challenges.

## Publications

### Peer-Reviewed Conference

#### Quantifying Misalignment Between Agents: Towards a Sociotechnical Understanding of Alignment

Aidan Kierans, [Avijit Ghosh](#), Hananel Hazan, Shiri Dori-Hacohen

AAAI '25

Philadelphia, USA

#### Coordinated Disclosure for AI: Beyond Security Vulnerabilities

Sven Cattell, [Avijit Ghosh](#), Lucie-Aimée Kaffee

AIES '24

San Jose, USA

#### Perceptions in pixels: analyzing perceived gender and skin tone in real-world image search results

Jeffrey Gleason, [Avijit Ghosh](#), Christo Wilson

WWW '24

Singapore

#### Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms

N. Dennler, A. Ovalle, A. Singh, L. Soldaini, A. Subramonian, H. Tu, W. Agnew, [Avijit Ghosh](#), K. Yee, I.F. Peradejordi, Z. Talat, M. Russo, J. Pinhal

AIES '23

Montreal, Canada

#### When Fair Classification Meets Noisy Protected Attributes

[Avijit Ghosh](#), Pablo Kvitca, Christo Wilson

AIES '23

Montreal, Canada

#### Queer In AI: A Case Study in Community-Led Participatory AI

Organizers of Queer In AI (50 authors)

FAccT '23

Chicago, Illinois

#### Subverting Fair Image Search with Generative Adversarial Perturbations

[Avijit Ghosh](#), Matthew Jagielski, Christo Wilson

FAccT '22

Seoul, South Korea

#### FairCanary: Rapid Continuous Explainable Fairness

[Avijit Ghosh\\*](#), Aalok Shanbhag\*, Christo Wilson

AIES '22

Oxford, United Kingdom

#### Algorithms that “Don’t See Color”: Comparing Biases in Lookalike and Special Ad Audiences

Piotr Sapiezynski, [Avijit Ghosh](#), Levi Kaplan, Alan Mislove, Aaron Rieke

AIES '22

Oxford, United Kingdom

#### When Fair Ranking Meets Uncertain Inference

[Avijit Ghosh](#), Ritam Dutt, Christo Wilson

SIGIR '21

Montreal, Canada / Virtual

#### Building and Auditing Fair Algorithms: A Case Study in Candidate Screening

Christo Wilson, [Avijit Ghosh](#), Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, Frida Polli

FAccT '21

Toronto, Canada / Virtual

#### Public Sphere 2.0: Targeted Commenting in Online News Media

Ankan Mullick, Sayan Ghosh\*, Ritam Dutt\*, [Avijit Ghosh\\*](#), Abhijnan Chakrabarty

ECIR '19

Cologne, Germany

## Peer-Reviewed Workshop

### To Err is AI: A Case Study Informing LLM Flaw Reporting Practices

Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, Will Smith, Shayne Longpre, Avijit Ghosh, Christopher Fiorelli, Michelle Hoang, Sven Cattell, Nouha Dziri

IAAI@AAAI '25

Philadelphia, USA

### Can There be Art Without an Artist?

Avijit Ghosh, Genoveva Fossas

HEGM@NeurIPS '22

New Orleans, USA

### Characterizing Intersectional Group Fairness with Worst-Case Comparisons

Avijit Ghosh, Lea Genuit, Mary Reagan

AIDBEI@AAAI '21

Vancouver, Canada / Virtual

### Analyzing Political Advertisers' Use of Facebook's Targeting Features

Avijit Ghosh, Giridhari Venkatadri, Alan Mislove

Conpro@S&P '19

San Francisco, USA

### SAVITR: A System for Real-time Location Extraction from Microblogs during Emergencies

Ritam Dutt, Kaustubh Hiware, Avijit Ghosh, Rameshwar Bhaskaran

SMERP@WWW '18

Lyon, France

### WebSelect: A Research Prototype for Optimizing Ad Exposures based on Network Structure

Avijit Ghosh, Agam Gupta, Divya Sharma, Uttam Sarkar

WITS'19

Dublin, Ireland

## Peer-Reviewed Journal

### Connectedness of Markets with Heterogeneous Agents and the Information Cascades

Avijit Ghosh, Aditya Chourasiya, Lakshay Bansal, Abhijeet Chandra

AAA'21

Journal

## Book Chapter

### Evaluating the social impact of generative AI systems in systems and society

Irene Solaiman, Zeerak Talat, et al. (including Avijit Ghosh)

Chapter

Oxford Handbook on Generative AI (forthcoming)

## Preprints and Working Manuscripts

### Dual Governance: The intersection of centralized regulation and crowdsourced safety mechanisms for Generative AI

Avijit Ghosh, Dhanya Lakshmi

Preprint

### Unified Shapley Framework to Explain Prediction Drift

Aalok Shanbhag\*, Avijit Ghosh\*, Josh Rubin\*

Preprint

### Supervised extraction of catchphrases from legal documents

Avijit Ghosh\*, Prerit Gupta\*, Ritam Dutt, Kaustubh Hiware, Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh

Term paper

\* Equal contribution

## Awards and Grants

2024	<b>Winner</b> Best AI Art	CVPR'24
2023	<b>Winner</b> Best Paper	FACCT'23
2022	<b>Winner</b> Best Paper - Runner Up	Conpro'22
2019	<b>Winner</b> Best Poster Award	ECIR'19
2019	<b>Dean's Fellowship</b> for first Year PhD students (USD 72K)	Northeastern University
2019	<b>Winner</b> Institute Order of Merit - Technology	IIT Kharagpur
2018	<b>Winner</b> SGSIS Institute Challenge Grant (INR 1M)	IIT Kharagpur
2017	<b>Silver Medal</b> Stock Market Analysis	Inter IIT Tech Meet, Kanpur
2016	<b>Gold Medal</b> Software Development	Inter IIT Tech Meet, Mandi
2012	<b>Governor's Medal</b> National Rank 5, ICSE Board	Government of West Bengal
2010	<b>NTSE Scholar</b> National Talent Search Examination	NCERT

## Teaching

### Responsible Machine Learning

Lecturer

Northeastern University

Fall 2023

- Explores the ethical challenges and responsibilities of creating and deploying machine learning (ML) models.
- Biases in ML models, methods for uncovering them, and algorithmic fairness techniques to mitigate them
- Term project to apply algorithmic fairness to a real world scenario

### Algorithmic Auditing

Teaching Assistant for Dr. Piotr Sapiezynski

Northeastern University

Spring 2022

- Designing audits that measure the effects of interests and control noise sources
- Minimize potential harms of audits to all stakeholders
- Legal bounds of algorithm audits
- Beyond audits: potential harms that cannot be measured through audits

## Academic Experience

---

### University of Connecticut

Storrs, CT

*Associate Researcher at Connecticut Advanced Computing Center*

Mar 2024 – Present

- Collaborating with Prof. Shiri Dori-Hacohen in the Risk and Information Ecosystem Threats (RIET) Lab on research and mentoring PhD students and Undergraduate Researchers on AI Alignment related projects as a visiting/associate expert.

### Northeastern University

Boston, MA

*Lecturer at Khoury College of Computer Sciences*

Sep 2023 – Present

- Teaching CS 4973-05 Responsible Machine Learning to senior undergrads. With the help of readings, guest lectures and original course material, the focus of the course is to empower students to responsibly deploy ethical and fair machine learning models for societal benefit.

### Northeastern University

Boston, MA

*Research Assistant at Khoury College of Computer Sciences*

Sep 2019 – Present

- Analyzing Fair ranking systems and showing how they fail in the presence of noisy protected attribute data. Also investigated adversarial attacks on the guaranteed fairness of retrieval systems.
- A cooperative fairness audit of the recommendation algorithm of [PyMetrics](#), a talent matching software. [Press Release](#).
- Investigated Facebook's Special Audiences system for opportunity advertisements and showed that the audience creation algorithm was still biased against women, seniors and minorities.
- Analyzed the ad reach and spend information obtained from Facebook's ad transparency feature and the personal targeting dataset from Propublica's Facebook ad dataset and showed that advertisers with higher budgets use more privacy sensitive targeting techniques like PII or Lookalike audiences. Findings published and presented at [IEEE ConPro 2019](#).

### LIG, University of Grenoble Alps

Grenoble, France

*Visiting Researcher*

May 2019 – July 2019

- Study of how news companies promote different items on social media, investigating possible patterns of differential information spreading using both posts and ads.
- We also discovered and reported an exposed access token bug to [Facebook Bug Bounty](#).

### IIT Kharagpur

Kharagpur, India

*Undergraduate Researcher - Complex Networks Research Group*

2014 – 2019

- Automated Extraction of Catchwords from Legal Documents using a novel NER tagger to help categorize lengthy legal texts.
- Automatically position user comments against relevant news article paragraphs. Presented at [ECIR 2019](#).
- Savitr - A real-time location extraction system for disaster management using twitter. Presented at [WWW-SMERP 2018](#).
- Classification and Summarization of tweets during a disaster event, presented at [IBM Day 2016](#).

## Industry Experience

---

### Hugging Face

New York, NY

*Applied Policy Researcher*

Mar 2024 – Present

- The Applied Policy Researcher position at Hugging Face involves working within the Machine Learning and Society team to bridge the gap between regulatory and technical realms. The role centers on developing tools to audit ML biases and engaging in policy discussions to facilitate understanding between policymakers and developers, with a focus on democratizing access to advanced machine learning technology.
- Responsibilities include contributing to ongoing public governance efforts by evaluating the social impacts of technology, providing feedback on regulatory proposals, and collaborating on projects such as governance in the BigCode project and evaluating the social impact of generative AI systems.

### AdeptID

Boston, MA

*Research Data Scientist*

Jul 2023 – Feb 2024

- Work with the Data Science and Engineering teams and external policy experts to audit internal systems and ensure that AdeptID's ML models are fair and unbiased and adhere to regulations.
- Conduct original research on ML Fairness and investigate solutions to unanswered questions about potential sources of bias in an industrially deployed ML pipeline.

### Twitter

San Francisco, CA

*Research Intern*

Sep – Dec 2021 and Jun – Aug 2022

- Worked with the META (Machine Ethics, Transparency and Accountability) team at Twitter, to investigate the relationship between demography agnostic and demography dependent author impression fairness metrics at scale.
- Developed home timeline diversity metrics based on user feedback, to find balance between recommendation efficiency and fairness.

## Fiddler Labs

Research Intern

Palo Alto, CA

Oct 2020 – Apr 2021

- Explain distributional shifts in Machine Learning model outputs by unifying Shapley based methods.
  - Using optimal transport theory, proposed a threshold independent fairness metric that allows for real time explanations.
  - Worked with the product team and civil rights lawyers in the deployment of Fiddler's Machine Learning model fairness dashboard.
- Introduced and incorporated intersectional fairness metrics in the product.

## Xerox Research Centre

Research Intern

Bangalore, India

May 2017 – July 2017

- Implemented XTrack, a Smart Vehicle Tracking and Battery usage minimizing Algorithm, using BLE to relay GPS information.
- Proposed a method for Uber-like Surge Price Prediction using Spatio-Temporal techniques like the Neural Hawkes and Recurrent Marked Temporal Point Process. Awarded the title of [Best Internship Project](#).

## Google Summer of Code

GSoC Student at OpenMRS

Remote

Apr 2016 - Aug 2016

- Replaced the HTML XForms system used with native generated forms using the Forms REST Api in the android client of the Opensource Medical Record System. Added offline form saving. Configured Travis CI to automatically build and push the apk to the play store.
- Overall, contributed [100K lines of code](#) and became the top code contributor in the project repository.

## Academic Service

2025	<b>ACM Conference on Fairness, Accountability, and Transparency</b>	Registration Chair
2024	<b>EvalEval: Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI</b>	General Chair
2024	<b>Conference on Neural Information Processing Systems</b>	Program Committee
2024	<b>ICWSM: The International AAAI Conference on Web and Social Media</b>	Program Committee
2024	<b>The Web Conference</b>	Program Committee
2024	<b>ACM Conference on Fairness, Accountability, and Transparency</b>	Program Committee
2023	<b>FAccTRec: Workshop on Responsible Recommendation</b>	Program Committee
2023	<b>Conference on Neural Information Processing Systems</b>	Program Committee
2023	<b>AAAI/ACM Conference on AI, Ethics, and Society</b>	Program Committee
2023	<b>ACM Conference on Fairness, Accountability, and Transparency</b>	Program Committee
2023	<b>The Web Conference</b>	Program Committee
2022	<b>ACM Conference on Fairness, Accountability, and Transparency</b>	Program Committee
2022	<b>AAAI/ACM Conference on AI, Ethics, and Society</b>	Program Committee
2022	<b>Conference on Neural Information Processing Systems</b>	Program Committee
2022	<b>Conference on Empirical Methods in Natural Language Processing</b>	Program Committee
2021	<b>Conference on Neural Information Processing Systems</b>	Program Committee

## Outreach and Leadership

2023	<b>SIGIR DEI lunch with speaker panel on disability in computing</b>	Chair
2023	<b>FAccT CRAFT Workshop on an India-first Responsible AI research agenda</b>	Organizer
2022	<b>FAccT CRAFT Workshop on Humanitarian AI for the Global South</b>	Organizer
2022	<b>FAccT CRAFT Workshop on Identifying Queer Harms as a bias bounty with Queer in AI</b>	Organizer
2021	<b>SIGIR Queer in AI social with speaker panel on queer stereotypes in web search</b>	Organizer
2017	<b>Kharagpur Winter Of Code</b>	Founder
2016	<b>Kharagpur Open Source Society</b>	Founder

## Speaking Engagements

2024	<b>Teaching Responsible AI: Building with open source and Hugging Face</b>	Northeastern University
2024	<b>Coordinated Disclosure for AI: Beyond Security Vulnerabilities</b>	EQUAL lab at MILA
2024	<b>Technology Impact on Cybersecurity</b>	Boston University
2024	<b>AI Vulnerability Reporting Event in the US Congress</b>	Hackers on the Hill
2023	<b>Queer Bias Bounty: Considerations for community audits and lessons learnt</b>	Centre for Data Ethics and Innovation, UK
2023	<b>Can There be AI Art Without an Artist?</b>	South By Southwest (SXSW)
2023	<b>On the evolving tension between centralized regulation and decentralized development of Text-to-Image Models</b>	AIDBEI at AAAI
2022	<b>Proxies for bias monitoring: Ethics workshop</b>	Centre for Data Ethics and Innovation, UK
2022	<b>Subverting Fair Image Search with Generative Adversarial Perturbations</b> as part of the 'Celebrating Young Researchers' event	Trustworthy ML Initiative

## Media Mentions

---

2025	<b>What does OpenAI get from Stargate? A \$500 billion chance to build a whole new moat.</b>	<i>Business Insider</i>
2025	<b>OpenAI's Stargate may be tech's biggest gamble ever, but here's what's really at stake</b>	<i>Fortune</i>
2025	<b>AI will transform everything from daily life to businesses</b>	<i>Dainik Bhaskar Op-ed</i>
2024	<b>This Wearable AI Notetaker Will Transcribe Your Meetings — and Someday, Your Entire Life</b>	<i>Wired</i>
2024	<b>En Californie, la loi de protection contre l'IA amendée pour apaiser la Silicon Valley</b>	<i>Les Echos</i>
2024	<b>We finally have a definition for open-source AI</b>	<i>MIT Technology Review</i>
2024	<b>World's biggest hacker fest spotlights AI's soaring importance in the high-stakes cybersecurity war—and its vulnerability</b>	<i>Fortune</i>
2024	<b>AI full of prejudices both a challenge and an opportunity for India</b>	<i>Dainik Bhaskar Op-ed</i>
2024	<b>Common image search results are overwhelmingly white, a new study finds</b>	<i>Fast Company</i>
2024	<b>As AI tools get smarter, they're growing more covertly racist, experts find</b>	<i>The Guardian</i>
2024	<b>LLMs become more covertly racist with human intervention</b>	<i>MIT Technology Review</i>
2023	<b>How Can AI Affect LGBTQIA+ People In India? Three Indian-Origin Queer Researchers Explain</b>	<i>IndiaTimes</i>
2023	<b>He hacked AI chatbots to find flaws and vulnerabilities. Now Northeastern's Avijit Ghosh is writing a report on combating these problems</b>	<i>Northeastern Global News</i>
2023	<b>Radio Interview - AirTalk</b>	<i>LAIST/NPR</i>
2023	<b>Radio Interview - As It Happens</b>	<i>CBC Canada</i>
2023	<b>When Hackers Descended to Test A.I., They Found Flaws Aplenty</b>	<i>The New York Times</i>
2023	<b>From accessibility efforts to ethical concerns, here are our AI takeaways from SXSW content creators should consider</b>	<i>Passionfruit</i>
2021	<b>NYC aims to be first to rein in AI hiring tools</b>	<i>Associated Press</i>
2021	<b>Auditors are testing hiring algorithms for bias, but there's no easy fix</b>	<i>MIT Technology Review</i>
2021	<b>New York City Proposes Regulating Algorithms Used in Hiring</b>	<i>Wired</i>
2021	<b>Supporting Responsible Use of AI and Equitable Outcomes in Financial Services</b>	<i>The Federal Reserve</i>
2019	<b>Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement</b>	<i>Propublica</i>
2019	<b>Facebook Agreed Not to Let Its Ads Discriminate. But They Still Can.</b>	<i>Mother Jones</i>