

# 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 outline these problems in fair ML systems, with the aim to gain a more complete understanding of the issues involved and to be able to provide technical and policy recommendations to overcome their real world implementation challenges.

## Publications

### Peer-Reviewed Conference

#### Stop treating 'AGI' as the north-star goal of AI research

B. Blili-Hamelin, C. Graziul, L. Hancox-Li, H. Hazan, E. El-Mhamdi, **Avijit Ghosh**, K. Heller, J. Metcalf, F. Murai, E. Salvaggio, A. Smart, T. Snider, M. Tighanimine, T. Ringer, M. Mitchell, S. Dori-Hacohen

ICML '25

Vancouver, Canada

#### In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI

S. Longpre, K. Klyman, R. Appel, S. Kapoor, R. Bommasani, M. Sahar, S. McGregor, **Avijit Ghosh**, B. Blili-Hamelin, N. Butters, A. Nelson, A. Elazari, A. Sellars, C. Ellis, D. Sherrets, D. Song, H. Geiger, I. Cohen, L. McIlvenny, M. Srikumar, M. Jaycox, M. Anderljung, N. Johnson, N. Carlini, N. Mialhe, N. Marda, P. Henderson, R. Portnoff, R. Weiss, V. Westerhoff, Y. Jernite, R. Chowdhury, P. Liang, A. Narayanan.

ICML '25

Vancouver, Canada

#### "It's not a representation of me": Examining Accent Bias and Digital Exclusion in Synthetic AI Voice Services

Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, **Avijit Ghosh**

FAccT '25

Athens, Greece

#### 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, A. Ovalle, A. Subramonian, A. Singh, C. Voelcker, D. Sutherland, D. Locatelli, E. Breznik, F. Klubička, H. Yuan, H. J. H. Zhang, J. Shriram, K. Lehman, L. Soldaini, M. Sap, M. Deisenroth, M. Pacheco, M. Ryskina, M. Mundt, M. Agarwal, N. McLean, P. Xu, A. Pranav, R. Korpan, R. Ray, S. Mathew, S. Arora, S. John, T. Anand, V. Agrawal, W. Agnew, Y. Long, Z. Wang, Z. Talat, **Avijit Ghosh**, N. Dennler, M. Noseworthy, S. Jha, E. Baylor, A. Joshi, N. Bilenko, A. McNamara, R. Gontijo-Lopes, A. Markham, E. Dông, J. Kay, M. Saraswat, N. Vytla, L. Stark.

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

<b>When Fair Ranking Meets Uncertain Inference</b>	SIGIR '21
<b>Avijit Ghosh</b> , Ritam Dutt, Christo Wilson	Montreal, Canada / Virtual
<b>Building and Auditing Fair Algorithms: A Case Study in Candidate Screening</b>	FAccT '21
Christo Wilson, <b>Avijit Ghosh</b> , Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, Frida Polli	Toronto, Canada / Virtual
<b>Public Sphere 2.0: Targeted Commenting in Online News Media</b>	ECIR '19
Ankan Mullick, Sayan Ghosh*, Ritam Dutt*, <b>Avijit Ghosh*</b> , Abhijnan Chakrabarty	Cologne, Germany
<b>Peer-Reviewed Workshop</b>	
<b>Protecting Human Cognition in the Age of AI</b>	Tools4Thoughts@CHI '25
Anjali Singh, Karan Taneja, Zhitong Guan, <b>Avijit Ghosh</b>	Yokohama, Japan
<b>To Err is AI: A Case Study Informing LLM Flaw Reporting Practices</b>	IAAI@AAAI '25
Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, Will Smith, Shayne Longpre, <b>Avijit Ghosh</b> , Christopher Fiorelli, Michelle Hoang, Sven Cattell, Nouha Dziri	Philadelphia, USA
<b>Can There be Art Without an Artist?</b>	HEGM@NeurIPS '22
<b>Avijit Ghosh</b> , Genoveva Fossas	New Orleans, USA
<b>Characterizing Intersectional Group Fairness with Worst-Case Comparisons</b>	AIDBEI@AAAI '21
<b>Avijit Ghosh</b> , Lea Genuit, Mary Reagan	Vancouver, Canada / Virtual
<b>Analyzing Political Advertisers' Use of Facebook's Targeting Features</b>	Conpro@S&P '19
<b>Avijit Ghosh</b> , Giridhari Venkatadri, Alan Mislove	San Francisco, USA
<b>SAVITR: A System for Real-time Location Extraction from Microblogs during Emergencies</b>	SMERP@WWW '18
Ritam Dutt, Kaustubh Hiware, <b>Avijit Ghosh</b> , Rameshwar Bhaskaran	Lyon, France
<b>WebSelect: A Research Prototype for Optimizing Ad Exposures based on Network Structure</b>	WITS'19
<b>Avijit Ghosh</b> , Agam Gupta, Divya Sharma, Uttam Sarkar	Dublin, Ireland
<b>Peer-Reviewed Journal</b>	
<b>Connectedness of Markets with Heterogeneous Agents and the Information Cascades</b>	AAA'21
<b>Avijit Ghosh</b> , Aditya Chourasiya, Lakshay Bansal, Abhijeet Chandra	Journal
<b>Book Chapter</b>	
<b>Evaluating the social impact of generative AI systems in systems and society</b>	Chapter
Irene Solaiman, Zeerak Talat, et al. (including <b>Avijit Ghosh</b> )	Oxford Handbook on Generative AI (forthcoming)
<b>Preprints and Working Manuscripts</b>	
<b>Fully Autonomous AI Agents Should Not be Developed</b>	Preprint
Margaret Mitchell, <b>Avijit Ghosh</b> , Alexandra Sasha Luccioni, Giada Pistilli,	
<b>Dual Governance: The intersection of centralized regulation and crowdsourced safety mechanisms for Generative AI</b>	Preprint
<b>Avijit Ghosh</b> , Dhanya Lakshmi	
<b>Unified Shapley Framework to Explain Prediction Drift</b>	Preprint
Aalok Shanbhag*, <b>Avijit Ghosh*</b> , Josh Rubin*	
<b>Supervised extraction of catchphrases from legal documents</b>	Preprint
<b>Avijit Ghosh*</b> , Prerit Gupta*, Ritam Dutt, Kaustubh Hiware, Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh	
* Equal contribution	
<b>Grants</b>	
2025 <b>Beyond Polarization: Fostering Plurality in Large Language Model Design and Development. (\$6K)</b>	UCHI AI Seed Fund
2018 <b>SGSIS Institute Challenge Grant. (₹1M)</b>	IIT Kharagpur
<b>Awards</b>	
2025 <b>Winner</b> Spotlight Poster	ICML '25
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	ECIR '19
2019 <b>Dean's Fellowship</b> for first Year PhD students (\$ 72K)	Northeastern University
2019 <b>Winner</b> Institute Order of Merit - Technology	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

Northeastern University

*Lecturer*

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

Northeastern University

*Teaching Assistant for Dr. Piotr Sapiezynski*

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

Research Data Scientist

Boston, MA

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

Research Intern

San Francisco, CA

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.

## Outreach and Leadership

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
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

## Academic Service

2025	<b>Conference on Language Modeling</b>	Program Committee
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

## Speaking Engagements

2025	<b>Panel: “Is U.S. Policy Ready for Agentic AI?”</b>	Center for Data Innovation
2025	<b>Panel on Open Source AI Regulation at the Summit on State AI Legislation</b>	Digital Ethics Center, Yale University
2025	<b>NLP and Generative AI Panel</b>	HBS Tech Conference
2025	<b>Coordinated Disclosure for AI: Beyond Security Vulnerabilities</b>	The MIT Sloan AI Conference
2024	<b>Coordinated Disclosure for AI: Beyond Security Vulnerabilities</b>	The Future of Third-Party AI Evaluation Workshop
2024	<b>Bridging the Gap: Real-World AI Biases and Responsible Governance Frameworks</b>	IIT Kharagpur
2024	<b>ERASED BY AI: Personal Experiences of the Psychological Impact of AI Bias</b>	Arthur AI Fest
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 as part of the 'Celebrating Young Researchers' event</b>	Trustworthy ML Initiative

## Media Mentions

2025	<b>California finally beats Big Tech in court</b>	Politico
2025	<b>Hugging-face experts warn: We shouldn't give AI agents full control</b>	T3N (German)
2025	<b>Why handing over total control to AI agents would be a huge mistake</b>	MIT Technology Review Op-Ed
2025	<b>The hidden cost of brainstorming with ChatGPT</b>	Business Insider
2025	<b>DeepSeek pioneers a new way for AI to 'reason'</b>	Science News Explores
2025	<b>Two founders built a jobs board for AI agents. Humans need not apply — but their skills are still required.</b>	Business Insider
2025	<b>Why truly open-source AI remains out of reach</b>	HT TechCircle
2025	<b>What does OpenAI get from Stargate? A \$500 billion chance to build a whole new moat.</b>	Business Insider
2025	<b>OpenAI's Stargate may be tech's biggest gamble ever, but here's what's really at stake</b>	Fortune
2025	<b>AI will transform everything from daily life to businesses</b>	Dainik Bhaskar Op-ed (Hindi)
2024	<b>This Wearable AI Notetaker Will Transcribe Your Meetings — and Someday, Your Entire Life</b>	Wired
2024	<b>In California, the controversial AI protection law is about to be passed</b>	Les Echos (French)
2024	<b>We finally have a definition for open-source AI</b>	MIT Technology Review
2024	<b>World's biggest hacker fest spotlights AI's soaring importance in the high-stakes cybersecurity war—and its vulnerability</b>	Fortune
2024	<b>AI full of prejudices both a challenge and an opportunity for India</b>	Dainik Bhaskar Op-ed (Hindi)
2024	<b>Common image search results are overwhelmingly white, a new study finds</b>	Fast Company
2024	<b>As AI tools get smarter, they're growing more covertly racist, experts find</b>	The Guardian
2024	<b>LLMs become more covertly racist with human intervention</b>	MIT Technology Review
2023	<b>How Can AI Affect LGBTQIA+ People In India? Three Indian-Origin Queer Researchers Explain</b>	IndiaTimes
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>	Northeastern Global News
2023	<b>Radio Interview - AirTalk</b>	LAIST/NPR
2023	<b>Radio Interview - As It Happens</b>	CBC Canada
2023	<b>When Hackers Descended to Test A.I., They Found Flaws Aplenty</b>	The New York Times
2023	<b>From accessibility efforts to ethical concerns, here are our AI takeaways from SXSW content creators should consider</b>	Passionfruit
2021	<b>NYC aims to be first to rein in AI hiring tools</b>	Associated Press
2021	<b>Auditors are testing hiring algorithms for bias, but there's no easy fix</b>	MIT Technology Review
2021	<b>New York City Proposes Regulating Algorithms Used in Hiring</b>	Wired
2021	<b>Supporting Responsible Use of AI and Equitable Outcomes in Financial Services</b>	The Federal Reserve

2019 **Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement**

*ProPublica*

2019 **Facebook Agreed Not to Let Its Ads Discriminate. But They Still Can.**

*Mother Jones*