Avijit Ghosh

Technical AI Policy Researcher, Hugging Face

(+1) 857-337-0180 | avijit@huggingface.co (Work) | avijitg22@gmail.com (Personal) | evijit.io | evijit | evijit

Responsible AI | Machine Learning | AI Policy | AI for Science | AI Governance

Education

Northeastern University — Boston, MA
Ph.D. in Computer Science (Advised by Dr. Christo Wilson) — *2019 - 2023*

Indian Institute of Technology (IIT) Kharagpur — Kharagpur, India
B.Tech. in Chemical Engineering, M.Tech in Financial Engineering, Minor in Computer Science — *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

Documenting Patterns of Exoticism of Marginalized Populations within Text-to-Image Generators — AIES '25
Sourojit Ghosh, Sanjana Gautam, Pranav Venkit, **Avijit Ghosh** — *Madrid, Spain*

Stop treating 'AGI' as the north-star goal of AI research — ICML '25
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* — *Vancouver, Canada*

In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI — ICML '25
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. Miailhe, N. Marda, P. Henderson, R. Portnoff, R. Weiss, V. Westerhoff, Y. Jernite, R. Chowdhury, P. Liang, A. Narayanan.* — *Vancouver, Canada*

"It's not a representation of me": Examining Accent Bias and Digital Exclusion in Synthetic AI Voice Services — FAccT '25
Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, **Avijit Ghosh** — *Athens, Greece*

Quantifying Misalignment Between Agents: Towards a Sociotechnical Understanding of Alignment — AAAI '25
Aidan Kierans, **Avijit Ghosh**, *Hananel Hazan, Shiri Dori-Hacohen* — *Philadelphia, USA*

Coordinated Disclosure for AI: Beyond Security Vulnerabilities — AIES '24
Sven Cattell, **Avijit Ghosh**, *Lucie-Aimée Kaffee* — *San Jose, USA*

Perceptions in pixels: analyzing perceived gender and skin tone in real-world image search results — WWW '24
Jeffrey Gleason, **Avijit Ghosh**, *Christo Wilson* — *Singapore*

Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms — AIES '23
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* — *Montreal, Canada*

When Fair Classification Meets Noisy Protected Attributes — AIES '23
Avijit Ghosh, *Pablo Kvitca, Christo Wilson* — *Montreal, Canada*

Queer In AI: A Case Study in Community-Led Participatory AI — FAccT '23
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. Dòng, J. Kay, M. Saraswat, N. Vytla, L. Stark.* — *Chicago, Illinois*

Subverting Fair Image Search with Generative Adversarial Perturbations — FAccT '22
Avijit Ghosh, *Matthew Jagielski, Christo Wilson* — *Seoul, South Korea*

FairCanary: Rapid Continuous Explainable Fairness — AIES '22
Avijit Ghosh*, *Aalok Shanbhag*, Christo Wilson* — *Oxford, United Kingdom*

Algorithms that "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences AIES '22
Piotr Sapiezynski, **Avijit Ghosh**, *Levi Kaplan, Alan Mislove, Aaron Rieke* *Oxford, United Kingdom*

When Fair Ranking Meets Uncertain Inference SIGIR '21
Avijit Ghosh, *Ritam Dutt, Christo Wilson* *Montreal, Canada / Virtual*

Building and Auditing Fair Algorithms: A Case Study in Candidate Screening FAccT '21
Christo Wilson, **Avijit Ghosh**, *Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, Frida Polli* *Toronto, Canada / Virtual*

Public Sphere 2.0: Targeted Commenting in Online News Media ECIR '19
Ankan Mullick, Sayan Ghosh, Ritam Dutt*,* **Avijit Ghosh***, *Abhijnan Chakrabarty* *Cologne, Germany*

Peer-Reviewed Workshop

Protecting Human Cognition in the Age of AI Tools4Thoughts@CHI '25
Anjali Singh, Karan Taneja, Zhitong Guan, **Avijit Ghosh** *Yokohama, Japan*

To Err is AI: A Case Study Informing LLM Flaw Reporting Practices IAAI@AAAI '25
Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, Will Smith, Shayne Longpre, *Philadelphia, USA*
Avijit Ghosh, *Christopher Fiorelli, Michelle Hoang, Sven Cattell, Nouha Dziri*

Can There be Art Without an Artist? HEGM@NeurIPS '22
Avijit Ghosh, *Genoveva Fossas* *New Orleans, USA*

Characterizing Intersectional Group Fairness with Worst-Case Comparisons AIDBEI@AAAI '21
Avijit Ghosh, *Lea Genuit, Mary Reagan* *Vancouver, Canada / Virtual*

Analyzing Political Advertisers' Use of Facebook's Targeting Features Conpro@S&P '19
Avijit Ghosh, *Giridhari Venkatadri, Alan Mislove* *San Francisco, USA*

SAVITR: A System for Real-time Location Extraction from Microblogs during Emergencies SMERP@WWW '18
Ritam Dutt, Kaustubh Hiware, **Avijit Ghosh**, *Rameshwar Bhaskaran* *Lyon, France*

WebSelect: A Research Prototype for Optimizing Ad Exposures based on Network Structure WITS'19
Avijit Ghosh, *Agam Gupta, Divya Sharma, Uttam Sarkar* *Dublin, Ireland*

Peer-Reviewed Journal

Connectedness of Markets with Heterogeneous Agents and the Information Cascades AAA'21
Avijit Ghosh, *Aditya Chourasiya, Lakshay Bansal, Abhijeet Chandra* *Journal*

Book Chapter

Evaluating the social impact of generative AI systems in systems and society Chapter
I. Solaiman, Z. Talat, W. Agnew, L. Ahmad, D. Baker, S. Blodgett, C. Chen, H. Daumé, *Oxford Handbook on Generative AI (forthcoming)*
J. Dodge, I. Duan, E. Evans, F. Friedrich, **Avijit Ghosh**, *U. Gohar, S. Hooker, Y. Jernite, R. Kalluri, A. Lusoli, A. Leidinger, M. Lin, X. Lin, S. Luccioni, J. Mickel, M. Mitchell, J. Newman, A. Ovalle, M. Png, S. Singh, A. Strait, L. Struppek, A. Subramonian.*

Preprints and Working Manuscripts

AI For Scientific Discovery is a Social Problem Preprint
Georgia Channing,* **Avijit Ghosh***

Fully Autonomous AI Agents Should Not be Developed Preprint
Margaret Mitchell, **Avijit Ghosh**, *Alexandra Sasha Luccioni, Giada Pistilli,*

Dual Governance: The intersection of centralized regulation and crowdsourced safety mechanisms for Generative AI Preprint
Avijit Ghosh, *Dhanya Lakshmi*

Unified Shapley Framework to Explain Prediction Drift Preprint
Aalok Shanbhag,* **Avijit Ghosh***, *Josh Rubin**

Supervised extraction of catchphrases from legal documents Preprint
Avijit Ghosh*, *Prerit Gupta*, Ritam Dutt, Kaustubh Hiware, Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh*

** Equal contribution*

Grants

2025 **Beyond Polarization: Fostering Plurality in Large Language Model Design and Development. ($6K)** *UCHI AI Seed Fund*
2018 **SGSIS Institute Challenge Grant. (₹1M)** *IIT Kharagpur*

Awards

2025	**Winner** Spotlight Poster	*ICML '25*
2024	**Winner** Best AI Art	*CVPR '24*
2023	**Winner** Best Paper	*FAccT '23*
2022	**Winner** Best Paper - Runner Up	*Conpro '22*
2019	**Winner** Best Poster	*ECIR '19*
2019	**Dean's Fellowship** for first Year PhD students ($ 72K)	*Northeastern University*
2019	**Winner** Institute Order of Merit - Technology	*IIT Kharagpur*
2017	**Silver Medal** Stock Market Analysis	*Inter IIT Tech Meet, Kanpur*
2016	**Gold Medal** Software Development	*Inter IIT Tech Meet, Mandi*
2012	**Governor's Medal** National Rank 5, ICSE Board	*Government of West Bengal*
2010	**NTSE Scholar** 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
Technical AI Policy Researcher — *Mar 2024 – Present*

- I am responsible for policy responses from Hugging Face to ongoing legislative and regulatory movements in AI, while my technical research examines critical challenges in AI safety: from algorithmic bias to agent autonomy to standardization efforts in AI Evaluation and Vulnerability Disclosure. I translate this expertise to varied stakeholders through publications at academic machine learning conferences and op-eds in media publications. My work has influenced AI regulation, established best practices for AI Documentation, and advanced the democratization of machine learning technology while ensuring it serves human wellbeing.

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 — Palo Alto, CA
Research Intern — *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 — Bangalore, India
Research Intern — *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 — Remote
GSoC Student at OpenMRS — *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	**ACM Conference on Fairness, Accountability, and Transparency**	*Registration Chair*
2024	**EvalEval: Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI**	*General Chair*
2023	**SIGIR DEI lunch with speaker panel on disability in computing**	*Chair*
2023	**FAccT CRAFT Workshop on an India-first Responsible AI research agenda**	*Organizer*
2022	**FAccT CRAFT Workshop on Humanitarian AI for the Global South**	*Organizer*
2022	**FAccT CRAFT Workshop on Identifying Queer Harms as a bias bounty with Queer in AI**	*Organizer*
2021	**SIGIR Queer in AI social with speaker panel on queer stereotypes in web search**	*Organizer*
2017	**Kharagpur Winter Of Code**	*Founder*
2016	**Kharagpur Open Source Society**	*Founder*

Academic Service

2025	**Conference on Language Modeling**	*Program Committee*
2024	**Conference on Neural Information Processing Systems**	*Program Committee*
2024	**ICWSM: The International AAAI Conference on Web and Social Media**	*Program Committee*
2024	**The Web Conference**	*Program Committee*
2024	**ACM Conference on Fairness, Accountability, and Transparency**	*Program Committee*
2023	**FAccTRec: Workshop on Responsible Recommendation**	*Program Committee*
2023	**Conference on Neural Information Processing Systems**	*Program Committee*
2023	**AAAI/ACM Conference on AI, Ethics, and Society**	*Program Committee*
2023	**ACM Conference on Fairness, Accountability, and Transparency**	*Program Committee*
2023	**The Web Conference**	*Program Committee*
2022	**ACM Conference on Fairness, Accountability, and Transparency**	*Program Committee*
2022	**AAAI/ACM Conference on AI, Ethics, and Society**	*Program Committee*
2022	**Conference on Neural Information Processing Systems**	*Program Committee*
2022	**Conference on Empirical Methods in Natural Language Processing**	*Program Committee*
2021	**Conference on Neural Information Processing Systems**	*Program Committee*

Speaking Engagements

2025	**Panel: "Is U.S. Policy Ready for Agentic AI?"**	*Center for Data Innovation*
2025	**Panel on Open Source AI Regulation at the Summit on State AI Legislation**	*Digital Ethics Center, Yale University*
2025	**NLP and Generative AI Panel**	*HBS Tech Conference*
2025	**Coordinated Disclosure for AI: Beyond Security Vulnerabilities**	*The MIT Sloan AI Conference*
2024	**Coordinated Disclosure for AI: Beyond Security Vulnerabilities**	*The Future of Third-Party AI Evaluation Workshop*
2024	**Bridging the Gap: Real-World AI Biases and Responsible Governance Frameworks**	*IIT Kharagpur*
2024	**ERASED BY AI: Personal Experiences of the Psychological Impact of AI Bias**	*Arthur AI Fest*
2024	**Teaching Responsible AI: Building with open source and Hugging Face**	*Northeastern University*
2024	**Coordinated Disclosure for AI: Beyond Security Vulnerabilities**	*EQUAL lab at MILA*
2024	**Technology Impact on Cybersecurity**	*Boston University*
2024	**AI Vulnerability Reporting Event in the US Congress**	*Hackers on the Hill*
2023	**Queer Bias Bounty: Considerations for community audits and lessons learnt**	*Centre for Data Ethics and Innovation, UK*
2023	**Can There be AI Art Without an Artist?**	*South By Southwest (SXSW)*
2023	**On the evolving tension between centralized regulation and decentralized development of Text-to-Image Models**	*AIDBEI at AAAI*
2022	**Proxies for bias monitoring: Ethics workshop**	*Centre for Data Ethics and Innovation, UK*
2022	**Subverting Fair Image Search with Generative Adversarial Perturbations** as part of the 'Celebrating Young Researchers' event	*Trustworthy ML Initiative*

Media Mentions

2025	**California finally beats Big Tech in court**	*Politico*
2025	**Hugging-face experts warn: We shouldn't give AI agents full control**	*T3N (German)*
2025	**Why handing over total control to AI agents would be a huge mistake**	*MIT Technology Review Op-Ed*
2025	**The hidden cost of brainstorming with ChatGPT**	*Business Insider*
2025	**DeepSeek pioneers a new way for AI to 'reason'**	*Science News Explores*
2025	**Two founders built a jobs board for AI agents. Humans need not apply — but their skills are still required.**	*Business Insider*
2025	**Why truly open-source AI remains out of reach**	*HT TechCircle*
2025	**What does OpenAI get from Stargate? A $500 billion chance to build a whole new moat.**	*Business Insider*
2025	**OpenAI's Stargate may be tech's biggest gamble ever, but here's what's really at stake**	*Fortune*
2025	**AI will transform everything from daily life to businesses**	*Dainik Bhaskar Op-ed (Hindi)*
2024	**This Wearable AI Notetaker Will Transcribe Your Meetings — and Someday, Your Entire Life**	*Wired*
2024	**In California, the controversial AI protection law is about to be passed**	*Les Echos (French)*
2024	**We finally have a definition for open-source AI**	*MIT Technology Review*
2024	**World's biggest hacker fest spotlights AI's soaring importance in the high-stakes cybersecurity war—and its vulnerability**	*Fortune*

2024	**AI full of prejudices both a challenge and an opportunity for India**	*Dainik Bhaskar Op-ed (Hindi)*
2024	**Common image search results are overwhelmingly white, a new study finds**	*Fast Company*
2024	**As AI tools get smarter, they're growing more covertly racist, experts find**	*The Guardian*
2024	**LLMs become more covertly racist with human intervention**	*MIT Technology Review*
2023	**How Can AI Affect LGBTQIA+ People In India? Three Indian-Origin Queer Researchers Explain**	*IndiaTimes*
2023	**He hacked AI chatbots to find flaws and vulnerabilities. Now Northeastern's Avijit Ghosh is writing a report on combating these problems**	*Northeastern Global News*
2023	**Radio Interview - AirTalk**	*LAIST/NPR*
2023	**Radio Interview - As It Happens**	*CBC Canada*
2023	**When Hackers Descended to Test A.I., They Found Flaws Aplenty**	*The New York Times*
2023	**From accessibility efforts to ethical concerns, here are our AI takeaways from SXSW content creators should consider**	*Passionfruit*
2021	**NYC aims to be first to rein in AI hiring tools**	*Associated Press*
2021	**Auditors are testing hiring algorithms for bias, but there's no easy fix**	*MIT Technology Review*
2021	**New York City Proposes Regulating Algorithms Used in Hiring**	*Wired*
2021	**Supporting Responsible Use of AI and Equitable Outcomes in Financial Services**	*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*