Inioluwa Deborah Raji

(+01) 925-858-4491
rajiinio@berkeley.edu
https://rajiinio.github.io

Education

Electrical Engineering and Computer Science, University of California, Berkeley Berkeley, U.S.A.
Ph.D. in Computer Science 2021 - 2026 *(expected)*
- Advisor: Prof. Benjamin Recht

Engineering Science, University of Toronto Toronto, Canada
Bachelors of Applied Science in Engineering Science (Robotics Engineering) 2014 - 2019

Notable Awards and Honors

- **Tech for Humanity Prize**, New America 2024
- **Rise25 Award**, Mozilla Foundation 2024
- **Time 100 Most Influential People in AI**, Time magazine 2023
- **AI 100**, Business Insider 2023
- **Forbes 30 Under 30**, Forbes 2021
- **Pioneer/ Barlow Award**, Electronic Frontier Foundation (EFF) 2020
- **35 Under 35 Innovator Award**, MIT Tech Review 2020
- **AI Innovation Award for "AI for Good"**, Venture Beat 2019
- **Best Paper Awards:**
  - IEEE Conference on Secure and Trustworthy Machine Learning (SatML)(2024)
  - ML-Retrospectives, Surveys & Meta-Analyses (ML-RSA) workshop at NeurIPS (2020)
  - AAAI/ACM AI, Ethics & Society (AIES) conference (2019)

Fellowships and Grants

- **UK AI Security Institute Challenge Fund**, UK Government (£100,000) 2025
- **Technology Public Interest X-Grant**, MacArthur Foundation ($500,000) 2022
- **Senior Technology Fellowship**, Mozilla Foundation ($120,000 annually) 2021-2025

Advisory Roles

AI Advisory Working Group
Federation of American Scientists (FAS) 2025 - Present

Safe and Secure AI Advisory Group
Canadian Artificial Intelligence Safety Institute(CAISI) 2025 - Present

SAIGE Council
Partnership on AI (PAI) 2025 - Present

OECD AI Advisory Expert Network
OECD AI Policy Observatory 2024 - Present

Network of Experts for UNSG's AI Advisory Body
United Nations High-Level Advisory Body on Artificial Intelligence 2024 - Present

AI Policy and Governance Working Group
Science, Technology, and Social Values (ST&SV) Lab, Institute of Advanced Study 2023 - Present

Center for Civil Rights and Technology Advisory Council
Leadership Conference on Civil and Human Rights, a civil society coalition including ACLU, AFL-CIO, and NAACP Legal Defense Fund, amongst others 2023 - Present

TeachAI Advisory Committee
Practitioner network led by Code.org, ETS, ISTE, and Khan Academy 2023 - Present

CDT AI Governance Lab Advisory Committee
Center for Democracy and Technology AI Governance Lab 2023 - Present

RealML Advisory Committee
Practitioner network of AI accountability advocates 2023 - Present

Health AI Partnership (HAIP) Advisory Council
Practitioner network led by Duke Health and Mayo Clinic 2023 - Present

Health AI Partnership (HAIP) Leadership Council
Practitioner network led by Duke Health and Mayo Clinic 2022 - 2023

Peer-Reviewed Publications

Full list of publications is available on my Google Scholar page.

Key research themes: *AI functionality in deployment; AI auditing; AI accountability; Responsible AI; AI governance; Machine learning evaluation methods; Institutional design of AI policy ecosystem; AI product safety; Research ethics.*
$\underline{(\alpha \rightarrow \beta)}$ indicates author lists in alphabetical order by last name.

H-index = 28, i10index = 40+, citations = 10,000+.

Pre-prints:

1. Jessica Dai, **Inioluwa Deborah Raji**, Benjamin Recht, and Irene Y Chen. Aggregated Individual Reporting for Post-Deployment Evaluation. *arXiv preprint arXiv:2506.18133*, 2025
2. Victor Ojewale, **Inioluwa Deborah Raji**, and Suresh Venkatasubramanian. Multi-lingual functional evaluation for large language models. *arXiv preprint arXiv:2506.20793*, 2025
3. Lydia T Liu*, **Inioluwa Deborah Raji***, Angela Zhou*, Luke Guerdan, Jessica Hullman, Daniel Malinsky, Bryan Wilder, Simone Zhang, Hammaad Adam, Amanda Coston, et al. Bridging Prediction and Intervention Problems in Social Systems. *arXiv preprint arXiv:2507.05216*, 2025

Peer-reviewed conference papers:

[**AISTATS'25**] 1. **Inioluwa Deborah Raji** and Lydia T Liu. Evaluating Prediction-based Interventions with Human Decision Makers In Mind. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025

[**ICML'25**] 2. Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, **Inioluwa Deborah Raji**, and Travis Zack. Position: Medical Large Language Model Benchmarks Should Prioritize Construct Validity. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025 (Oral)

[**ICML'25**] 3. $\underline{(\alpha \rightarrow \beta)}$ Jessica Dai*, Paula Gradu*, **Inioluwa Deborah Raji***, and Benjamin Recht*. From individual experience to collective evidence: A reporting-based framework for identifying systemic harms. In *Forty-Second International Conference on Machine Learning*, 2025

[CHI '25] 4. Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and **Inioluwa Deborah Raji**. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2025

[FAccT '24] 5. Hellina Hailu Nigatu and **Inioluwa Deborah Raji**. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 141–160, 2024

[MLHC '24] 6. Judy Hanwen Shen*, **Inioluwa Deborah Raji***, and Irene Y Chen. The Data Addition Dilemma. *Proceedings of the 9th Machine Learning for Healthcare Conference*, 252, 2024

[SatML '24] 7. Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and **Inioluwa Deborah Raji**. SoK: AI Auditing: The Broken Bus on the Road to AI Accountability. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, 2024 (Best Paper Award)

[FAccT '23] 8. Jee Young Kim, William Boag, Freya Gulamali, Alifia Hasan, Henry David Jeffry Hogg, Mark Lifson, Deirdre Mulligan, Manesh Patel, **Inioluwa Deborah Raji**, Ajai Sehgal, et al. Organizational governance of emerging technologies: AI adoption in healthcare. In *proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1396–1417, 2023

[AIES '22] 9. **Inioluwa Deborah Raji**, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022

[FAccT '22] 10. Sasha Costanza-Chock, **Inioluwa Deborah Raji**, and Joy Buolamwini. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583, 2022

[FAccT '22] 11. **Inioluwa Deborah Raji**, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022

[FAccT '22] 12. Carolyn Ashurst, Solon Barocas, Rosie Campbell, and **Inioluwa Deborah Raji**. Disentangling the components of ethical research in machine learning. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2057–2068, 2022

[NeurIPS '21] 13. **Inioluwa Deborah Raji**, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021

[FAccT '21] 14. **Inioluwa Deborah Raji**, Morgan Klaus Scheuerman, and Razvan Amironesei. You can't sit with us: Exclusionary pedagogy in AI ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 515–525, 2021

[NeurIPS '21] 15. Thomas Liao, Rohan Taori, **Inioluwa Deborah Raji**, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021

[FAccT '20] 16. **Inioluwa Deborah Raji**, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, pages 33–44, 2020

[**AIES '20**] 17. **Inioluwa Deborah Raji**, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020

[**AIES '19**] 18. **Inioluwa Deborah Raji** and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019 (Best Paper Award)

[**FAccT '19**] 19. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, **Inioluwa Deborah Raji**, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019

Workshop papers:

[**AAAI '21**] 20. **Inioluwa Deborah Raji** and Genevieve Fried. About face: A survey of facial recognition evaluation. In *AAAI Workshop on AI Evaluation*, 2021

[**ICLR '20**] 21. **Inioluwa Deborah Raji** and Roel Dobbe. Concrete problems in AI safety, revisited. *ICLR workshop on ML in the Real World*, 2020

[**NeurIPS '19**] 22. Alice Xiang and **Inioluwa Deborah Raji**. On the legal compatibility of fairness definitions. In *NeurIPS Workshop on Human-Centric Machine Learning*, 2019

[**NeurIPS '19**] 23. **Inioluwa Deborah Raji** and Jingying Yang. ABOUT ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. In *NeurIPS Human-Centric Machine Learning Workshop*, 2019

Journal articles:

24. Mark P Sendak, Jee Young Kim, Alifia Hasan, Will Ratliff, Mark A Lifson, Manesh Patel, **Inioluwa Deborah Raji**, Ajai Sehgal, Keo Shaw, Danny Tobey, et al. Empowering US healthcare delivery organizations: Cultivating a community of practice to harness AI and advance health equity. *PLOS Digital Health*, 3(6):e0000513, 2024

25. Jee Young Kim, Alifia Hasan, Katherine C Kellogg, William Ratliff, Sara G Murray, Harini Suresh, Alexandra Valladares, Keo Shaw, Danny Tobey, David E Vidal, and **others**. Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities. *PLOS Digital Health*, 3(5):e0000390, 2024

26. Sayash Kapoor, Emily M Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A Bail, Odd Erik Gundersen, Jake M Hofman, Jessica Hullman, Michael A Lones, Momin M Malik, and **others**. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18):eadk3452, 2024

27. Neil Guha, Christie M Lawrence, Lindsey A Gailmard, Kit T Rodolfa, Faiz Surani, Rishi Bommasani, **Inioluwa Deborah Raji**, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *Geo. Wash. L. Rev.*, 92:1473, 2024

28. **Inioluwa Deborah Raji** and Joy Buolamwini. Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial ai products. *Communications of the ACM*, 66(1):101–108, 2022

29. Amandalynne Paullada, **Inioluwa Deborah Raji**, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Cell Patterns*, 2(11), 2021

Book Chapters

1. **Inioluwa Deborah Raji**. "The Anatomy of AI Audits: Form, Process, and Consequences." *The Oxford handbook of AI governance*, edited by Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, Baobao Zhang, Oxford University Press (2024).
2. **Inioluwa Deborah Raji**, Danaë Metaxa. "A Tale of Two Audits." *AI and Assembly: Coming Together and Apart in a Datafied World*, edited by Toussaint Nothias and Lucy Bernholz, Stanford University Press (2025).
3. **Inioluwa Deborah Raji**, Sasha Costanza Chock, Joy Buolamwini. "Change from the outside: towards credible third-party audits of AI systems." *Missing links in AI governance*, edited by Edited by Benjamin Prud'homme, Catherine Régis, Golnoosh Farnadi, Vanessa Dreier, Sasha Rubel, Charline d'Oultremont, United Nations Educational, Scientific and Cultural Organization (UNESCO) (2023).
4. **Inioluwa Deborah Raji**. "Facing the Tech Giants." *The Black Agenda: Bold Solutions for a Broken System*, edited by Anna Gifty Opoku-Agyeman, St. Martin's Publishing Group (2022).
5. **Inioluwa Deborah Raji**. "The bodies underneath the rubble." *Fake AI*, edited by Frederike Kaltheuner, Meatspace Press (2021).

Other Publications

Journal editorials & comments:

1. L Weidinger*, **Inioluwa Deborah Raji***, H Wallach, M Mitchell, A Wang, O Salaudeen, R Bommasani, D Ganguli, S Koyejo, and W Isaac. Toward an evaluation science for generative AI systems. *National Academy of Engineering, Spring Bridge on AI: Promises and Risks*
2. **Inioluwa Deborah Raji**, Roxana Daneshjou, and Emily Alsentzer. It's time to bench the medical exam benchmark. *New England Jornal of Medicine (NEJM) AI*, 2(2):AIe2401235, 2025
3. **Inioluwa Deborah Raji**. Handle with care: lessons for data science from black female scholars. *Cell Patterns*, 1(8), 2020
4. **Inioluwa Deborah Raji**. The discomfort of death counts: mourning through the distorted lens of reported COVID-19 death data. *Cell Patterns*, 1(4), 2020
5. **Inioluwa Deborah Raji**. That's not fair! *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):44–48, 2019

Popular Press Articles:

1. **Inioluwa Deborah Raji**. AI's Present Matters More Than Its Imagined Future. *The Atlantic*, 2023
2. Abeba Birhane and **Inioluwa Deborah Raji**. ChatGPT, Galactica, and the progress trap. *WIRED*, 2022
3. **Inioluwa Deborah Raji**. How our data encodes systematic racism. *MIT Technology Review*, 123(6), 2020

Full workshop proceedings:

1. **Inioluwa Deborah Raji**, Carolyn Ashurst, Solon Barocas, Rosie Campbell, and Stuart Russell. NeurIPS Workshop on Navigating the Broader Impacts of AI Research. In *Thirty-fourth Conference on Neural Information Processing Systems*, 2020
2. Bogdan Kulynych, David Madras, Smitha Milli, **Inioluwa Deborah Raji**, Angela Zhou, and Richard Zemel. Participatory approaches to machine learning. In *International Conference on Machine Learning Workshop*, volume 7, 2020

Other Publications

Public Policy Publications & Reports:

1. ($\alpha \rightarrow \beta$) Rishi Bommasani*, Scott R Singer*, Ruth E Appel, Sarah Cen, A Feder Cooper, Lindsey A Gailmard, Ian Klaus, Meredith M Lee, **Inioluwa Deborah Raji**, Anka Reuel, et al. The California Report on Frontier AI Policy. *arXiv preprint arXiv:2506.17303*, 2025 (commissioned by Governor Newsom as part of the Joint California Policy Working Group on AI Frontier Models; strongly informed the drafting and passing of AI California State Bill SB53)

2. ($\alpha \rightarrow \beta$) B* Yohsua, P* Daniel, B Tamay, B Rishi, C Stephen, C Yejin, G Danielle, H Hoda, K Leila, L Shayne, and **others**. International scientific report on the safety of advanced AI. *Department for Science, Innovation and Technology, Tech. Rep.*, 2024 (with UK Government, Department for Science, Innovation and Technology, distributed to government Digital Ministers worldwide at global AI Safety Summits)

3. ($\alpha \rightarrow \beta$) Jon Bateman, Dan Baer, Stephanie A Bell, Glenn O Brown, Mariano-Florentino Tino Cuéllar, Deep Ganguli, Peter Henderson, Brodi Kotila, Larry Lessig, Nicklas Berild Lundblad, and **others**. Beyond open vs. closed: Emerging consensus and key questions for foundation AI model governance. 2024 (with Carnegie Endowment for International Peace)

4. Josh Kenway, Camille François, Sasha Costanza-Chock, **Inioluwa Deborah Raji**, and Joy Buolamwini. Bug bounties for algorithmic harms. *Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress. Algorithmic Justice League, Washington, DC*, 2022

5. ($\alpha \rightarrow \beta$) Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, and **others**. AI Now 2019 Report. 2019

Blogs & Misc:

1. **Inioluwa Deborah Raji**. It's time to develop the tools we need to hold algorithms accountable. *Mozilla Foundation*, 2022

2. Sasha Luccioni, William Isaac, Cherie Poland, and **Inioluwa Deborah Raji**. Reflections on the NeurIPS 2022 Ethics Review Process. *Nerural Information Processing Systems Conference*, 2022

3. Sasha Luccioni, William Isaac, Cherie Poland, **Inioluwa Deborah Raji**, Samy Bengio, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, and Marc'Aurelio Ranzato. Ethical Review Guidelines. *Nerural Information Processing Systems Conference*, 2022

4. Samy Bengio, Alina Beygelzimer, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, **Inioluwa Deborah Raji**, and Marc'Aurelio Ranzato. Provisional draft of the NeurIPS code of ethics. *Nerural Information Processing Systems Conference*, 2022

5. Samy Bengio, **Inioluwa Deborah Raji**, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. A retrospective on the NeurIPS 2021 ethics review process. *Nerural Information Processing Systems Conference*, 2021

6. Samy Bengio, Alina Beygelzimer, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, **Inioluwa Deborah Raji**, and Marc'Aurelio Ranzato. NeurIPS 2021 Ethics Guidelines. *Nerural Information Processing Systems Conference*, 2021

Select Policy Engagements

Speaker, Western State Legislators Briefing
American Association for the Advancement of Science (AAAS) 2025

Public Interest AI Working Group, Paris AI Action Summit
French Government, Envoy to the AI Action Summit 2024 - 2025

Principal Participant, AI Academic Roundtable
Organized by United States House of Representative Ro Khanna 2024

Speaker, Conference on Frontier AI Safety Frameworks
Organized by U.K. AI Safety Institute 2024

Participant, Inaugural Convening of the International Network of AI Safety Institutes
Organized by International Network of AI Safety Institutes, chaired by U.S. AISI 2024

Panelist, Artificial Intelligence Hearing
Office of the National Coordinator for Health IT (ONC) 2024

Delegate, AI Safety Summit
U.K. Government, Department for Science, Innovation and Technology 2023

Principal Participant, Civil Rights Roundtable
Organized by United States Senator Cory Booker 2023

Speaker, Congressional Briefing on Artificial Intelligence and Civil Rights
Organized by Lawyer's Committee For Civil Rights 2023

Principal Participant, Inaugural AI Insight Forum
Organized by United States Senate Leader Chuck Shumer 2023

1. **Work cited in the following select government reports, letters and official communications:**
   - US National AI Advisory Committee (NAIAC) Report of Findings & Recommendations on AI Safety (2024)
   - Economic Report of the President(2024)
   - House Task Force on Artificial Intelligence Final Report(2024)
   - The Bipartisan Senate AI Working Group report(2024)
   - Letter from House Committee on House Committee on Science, Space, and Technology(2023)
   - National Institute of Standards and Technology (NIST) AI Risk Management Framework Playbook(2023)
   - Government Accountability Office (GAO) Accountability Framework for Federal Agencies and Other Entities(2021)
   - National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) Part 3 Demographic Effects (2019)

   and more.
2. **Led policy comment submissions with Mozilla Open Source Audit Tooling (OAT) team:**
   - US National Telecommunication and Information Administration (NTIA) Comment (cited several times in their following "Artificial Intelligence Accountability Policy" Report!)
   - US National Institute on Standards & Technology (NIST) Comment
   - EU Digital Service Act (DSA), Article 40 Comment (cited several times in report for Delegated Act on data access provided for in Digital Services Act)
3. **Participated** in policy comment submissions on open source, AI accountability and third party auditor access, with the Institute of Advanced Study Science, Technology & Social Values (ST&SV) Lab AI Policy and Governance Working Group, Stanford Human-centered AI (HAI), Carnegie Endowment for International Peace (CEIP), and Mozilla.

Select Media Features & Impact

My work has been widely featured in mainstream media regularly.

Select Popular Press Features: I have been quoted and have had my work featured in several articles for the New York Times, WIRED, Washington Post, MIT Technology Review, Nature, Venture Beat and more. Some key highlights include:

- "In Show of Force, Silicon Valley Titans Pledge 'Getting This Right' With A.I."(Cecilia Kang, New York Times, Sept. 13, 2023)
- "Meet the Humans Trying to Keep Us Safe From AI"(Will Knight, Khari Johnson, Morgan Meaker, WIRED, Jun 27, 2023)
- "Who Is Making Sure the A.I. Machines Aren't Racist?"(Cade Metz, New York Times, Mar 15 2021)
- "Is facial recognition too biased to be let loose?"(Davide Castelvecchi, Nature, Nov 18 2020)
- "The two-year fight to stop Amazon from selling face recognition to the police"(Karen Hao, MIT Tech Review, Jun 12 2020)
- "Google researchers release audit framework to close AI accountability gap"(Khari Johnson, VentureBeat, Jan 30 2020)
- "Amazon Is Pushing Facial Technology That a Study Says Could Be Biased"(Natasha Singer, New York Times, Jan 24 2019)

Select Film, Video & Podcast Features: I've been featured in radio interviews with BBC, CBC and podcasts such as the Public Book podcast and Slate's TBD podcast. I've been featured in videos for Code.org, Bloomberg, and Vox, and will appear in various upcoming full-length documentary films, including *Coded Bias (dir. Shalini Kantayya, 2020)*.

Select Book Features: Our work is regularly cited and featured in published books. Some notable examples include:

- Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World (Cade Metz, 2022)
- Your Face Belongs to Us: A Tale of AI, a Secretive Startup, and the End of Privacy (Kashmir Hill, 2024)
- Unmasking AI: My Mission to Protect What Is Human in a World of Machines (Joy Buolamwini, 2024)
- The AI Con: How to Fight Big Tech's Hype and Create the Future We Want (Dr. Emily M. Bender, Dr. Alex Hanna, 2025)
- Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI (Karen Hao, 2025)
- More Everything Forever: AI Overlords, Space Empires, and Silicon Valley's Crusade to Control the Fate of Humanity (Adam Becker, 2025)

Other: Model Cards are widely adopted in industry (used by Google, OpenAI, Anthropic, Cohere, Nvidia, xAI, ByteDance, Meta, Perspective API, Citadel AI etc.); research (recommended by NeurIPS ethical guidelines, ACL ARR Responsible NLP Research checklist, etc.); and policy (featured in OECD Catalogue of Tools & Metrics for Trustworthy AI, UNICEF Open Source Inventory, etc.). Also integrated into model development platforms such as scikit-learn, and into Hugging Face.

Both Model Cards and the SMACTR Framework have been widely adopted in healthcare - cited by both the Coalition for Health AI (CHAI), Health AI Partnership (HAIP), Office of the National Coordinator for Health Information Technology (ONC) and in the Congressional Research Service report on "Artificial Intelligence (AI) in Health Care". Beyond the United States, this work has also been cited extensively in an Ada Lovelace Institute report on behalf of the United Kingdom National Health Service (NHS) AI Lab, on the use of an algorithmic impact assessment for data access in a healthcare context.

Academic Service

Main Organizer for:

- Simons Institute workshop on "Bridging Prediction and Intervention Problems in Social Systems"[2026]
- Banff International Research Station (BIRS) for Mathematical Innovation and Discovery workshop on "Bridging Prediction and Intervention Problems in Social Systems" workshop [2024]
- SSRC Data Fluencies Research Workshop [2024]
- ICML Workshop on "Participatory Approaches to Machine Learning" [2020]
- NeurIPS Workshop on "Navigating the Broader Impacts of AI Research" [2020]
- REALML workshop [2019, 2021]
- **UC Berkeley Initiatives**
  - Algorithms, Data & Society (ADS) Reading Group [Weekly; 2022-2025]
  - AI & Society Initiative [Bi-weekly to monthly; 2024-present]

Organizing Committee:

- **Steering Committee**, Fairness, Accountability & Transparency (FAccT) Conference [2021, 2022, 2023, 2024, 2025]
- **Ethics Review Chair**, Neural Information Processing Systems (NeurIPS) Conference [2021, 2022]
- **Publicity co-chair**, Fairness, Accountability & Transparency Conference (FAccT) [2021, 2022]

Area Chair for:

- Fairness, Accountability & Transparency Conference (FAccT) [Data and Algorithm Evaluation; 2021-present]
- Neural Information Processing Systems (NeurIPS) Conference [Datasets & Benchmarks track; 2023-present]

Reviewer for:

- **Journals** - Nature, Cell Patterns, Big Data & Society
- **Conferences** - International Conference of Machine Learning (ICML), Neural Information Processing Systems (NeurIPS), ACM Fairness, Accountability & Transparency (FAccT) conference, AAAI/ACM Artificial Intelligence, Ethics, & Society (AIES) conference
- **Funders** - Mozilla Technology Fund (MTF), Canadian Institute for Advanced Research (CIFAR), Schmidt Sciences AI2050

Teaching & Course Development

PP290: "Case Studies in Prediction, AI, and Public Policy" | Co-instructor **Fall 2025**

- UC Berkeley, Goldman School of public policy course, open both to policy (e.g., MPP and MPA) students interested in data science as well as data science, EECS and statistics PhD students and advanced undergraduate students interested in policy.
- Developed full syllabus and teaching materials (teaching slides, reading assignments, course evaluation, class pre-read) ; will manage grading, and office hours.
- Invited and will manage roster of 7 guest speakers throughout the course.
- Will lecture regularly, with co-Instructor Prof. Avi Feller, about once a week.

CS281B/Stat241B: "Machine Learning Evaluation" | Co-instructor **Spring 2025**

- UC Berkeley, Electrical Engineering and Computer Science (EECS) graduate course for CS/Stats/Engineering Masters and PhD students.
- Developed full syllabus and teaching materials (teaching slides, reading assignments, course evaluation); managed grading, and office hours.
- Lectured regularly, with co-Instructor Prof. Benjamin Recht, about once a week.

Select Guest Lectures

My work has been featured in a variety of syllabi across a wide range of courses, and I am often invited to give guest lectures as a result. Here is a sample of recent guest lectures:

"CS120: Introduction to AI Safety"
Stanford University Fall 2025

"DATA 2030: Forces of Influence in AI Governance"
Brown University Fall 2025

"ORIE 5355/INFO 5370: Applied Data Science - Decision-making beyond Prediction"
Cornell University Fall 2025

AI Policy Summer School
Brown University Summer 2025

Computational Social Science Summer School
Stanford University Summer 2025

"6.S977: Ethical Machine Learning in Human Deployments"
Massachusetts Institute of Technology, EECS Spring 2025

"LTI 11801: Quantitative Evaluation of Language Technologies"
Carnegie Mellon University Spring 2025

"CS 269: Computational Ethics, Large Language Models and the Future of NLP"
University of California, Los Angeles Spring 2025

"CS 294-297: Data Science for Social Change"
University of California, Berkeley Fall 2024

"UGBA 192T: Responsible AI Innovation & Management"
UC Berkeley Haas School of Business Spring 2024

"CS 5382: Practical Principles for Designing Fair Algorithms"
Cornell University Spring 2024

Speaker, AI Cyber Lunch
Harvard Kennedy School Spring 2023

Speaker, Yale ISP Ideas Lunch
Yale Law School Spring 2023

"INTLPOL 323/LAW 7082: Free Speech, Democracy and the Internet"
Stanford Law School Fall 2022

"INAF U6202: Internet Governance and Human Rights"
Columbia University School of International and Public Affairs Spring 2022

"CS4910: Special Topics in Computer Science: Algorithm Audits"
Northeastern University Spring 2021

"SPRING POLICY LAB: AI and Implicit Bias"
University of Pennsylvania, Carey Law School Spring 2021

Select Named Lectures & Seminars

Distinguished Speaker Series
University of Colorado Boulder Fall 2025 (Upcoming)

William Pierson Field Lecture
Princeton University Spring 2024

Distinguished Speaker Series, Program in Criminal Justice Policy and Management
Harvard Kennedy School Spring 2024

Fred Kan Distinguished Lecture
University of Toronto Fall 2023

Speaker, AI.Humanity Seminar
Emory University Fall 2023

Speaker, Values-Centered Artificial Intelligence (VCAI) Colloquium
University of Maryland Fall 2023

Distinguished Speaker Seminar Series, Yale Center for Biomedical Data Science (CBDS)
Yale School of Medicine Spring 2021

Select Conference Keynotes, Panels & Invited Talks

Speaker, "Benchmarking AI Systems"
American Physical Society workshop on AI and the Practice of Physics 2025

Keynote, "Towards a sociotechnical view on AI Safety"
Canadian AI Safety Institute (CAISI) Research Program Annual Meeting 2025

Panelist, "The States Shaping AI's Fate"
Berkeley Public Policy Conference 2025

Panelist, "AI and geopolitical power"
Financial Times Future of AI Summit 2025

Keynote, "Understanding and Addressing Gender Bias in Artificial Intelligence Systems"
United Nations Women AI-Gender Academic Conference 2025

Panelist
Conference on Health, Inference, and Learning (CHIL) 2025

Panelist, "How To Put The Missing Human Values Back Into AI"
American Medical Informatics Association 2025

Speaker, "A Primer on Algorithm Bias"
GovAI Coalition Summit 2025

Keynote, "Safety, By Any Other Name"
Next Generation of AI Safety workshop, ICML 2024

Keynote
Conference on Health, Inference, and Learning (CHIL) 2024

Panelist
National Artificial Intelligence Advisory Committee (NAIAC) Panel on AI Safety 2024

Keynote
State Legislative Leaders Foundation Conference 2024

Keynote
2nd IEEE Conference on Secure and Trustworthy Machine Learning 2024

Select Conference Keynotes, Panels & Invited Talks

Panelist
Workshop on Regulating Machine Learning, NeurIPS 2023

Keynote, "Grounded Evaluations for Assessing Real-World Harms"
Socially Responsible Language Modelling Research (SoLaR) Workshop, NeurIPS 2023

Keynote
ACM/AAAI AI, Ethics and Society (AIES) Conference 2022

Keynote, "Ethical Challenges of Data Collection & Use in Machine Learning Research"
DataPerf: Benchmarking Data for Data-Centric AI, ICML 2022

Panel Moderator, "How Should a Machine Learning Researcher Think About AI Ethics?"
Neural Information Processing Systems (NeurIPS) 2021

Keynote, "Radical Proposal: Third-Party Auditor Access for AI Accountability"
2021 HAI Fall Conference on Policy & AI 2021

Keynote, "The Need for Ethical Oversight in Machine Learning Research"
Journal of Opportunities, Unexpected limitations, Retrospectives, Negative results, and Experiences, MLSys 2021

Keynote, "AI Resistance and the Five Stages of Corporate Grief"
Resistance AI Workshop, NeurIPS 2020

Keynote, "Classic examples of engineering responsibility"
MLRetrospectives: A Venue for Self-Reflection in ML Research workshop, ICML 2020

Keynote, "Debugging" Discriminatory ML Systems"
Debugging Machine Learning Models workshop, ICLR 2019

Work Experience

Founder, Open Source Audit Tooling (OAT) | Mozilla Foundation 2022 - 2025
- Recruited and led a team of five (3 researchers and 2 research assistants) to conduct research on the open source technologies needed to resource AI audit practice.
- Published two academic peer-reviewed papers, one of which won a Best Paper award. Submitted 3 policy comments, two of which were cited in follow up reports by the US NTIA and the European Commission; gave presentations to US Federal Trade Commission (FTC), UK Ofcom, and OECD.
- Advised selection of Mozilla Technology Fund (MTF) grantees: 5 projects funded in inaugural cohort (2022); 8 projects in follow-up cohort (2023) on AI audit tooling.

Trustworthy AI Research Fellow | Mozilla Foundation 2021-2023
- Independent researcher working on the topics of AI accountability and evaluation.

Graduate Fellow | Stanford RegLab June 2022- Present
- Working with legal collaborators on regulatory proposals for AI auditing.

Harms Research Fellow | Algorithmic Justice League Oct 2020 - Sept 2021
- Developing projects on designing reporting systems for harms discovery in algorithmic auditing.

Technology Fellow | AI Now Institute, New York University Oct 2019 - Oct 2020
- Research and policy work in facial recognition and AI accountability practice.

Research Fellow & Affiliate | Partnership on AI June 2019 - Jan 2020
- Setting transparency standards, benchmarking norms and legal conditions for AI deployment.

Student Researcher | Google AI June 2018 - Nov 2019
- Selected for 12 month Google AI Research Mentorship Program, paired with the Ethical AI team.