Renata Barreto
Sr. AI Research Scientist | Risk and Evaluation

San Francisco, California
210.639.3924 | rbarreto@berkeley.edu | GitHub: rbarreto

SUMMARY

AI evaluation researcher working on model behavior, safety, and alignment, with a sociotechnical approach grounded in computational methods and science and technology studies. Current research examines the durability of model behavior under fine-tuning and whether synthetic data can stand in for human experience, bridging empirical evaluation with questions about which behaviors and capabilities models should have and the values they embody. At eBay, owns model release readiness review and develops AI agent governance in partnership with engineering, product, legal, and compliance.

EDUCATION

UNIVERSITY OF CALIFORNIA – BERKELEY *2025*
Ph.D. in Jurisprudence and Social Policy (Computational Social Science Field Focus)
Dissertation: Artificial Intelligence and the Operationalization of SESTA-FOSTA on Social Media Platforms

Designated Emphasis in Science and Technology Studies
Graduate Certificate in Applied Data Science

J.D. in Anti-Discrimination Law & Technology *2022*
Law & Technology Certificate

REED COLLEGE, Portland, OR *2013*
B.A. in International & Comparative Policy Studies – Political Science & Economics

EXPERIENCE

SR. RESEARCH SCIENTIST, RESPONSIBLE AI EVALUATIONS, RISK, AND COMPLIANCE
eBay, San Jose, CA *June 2025 – present*

- Own model release readiness for 1,000+ annual AI reviews and author policies governing risk-tiered evaluation, AI agent autonomy, and human oversight.
- Partner across engineering, product, legal, and compliance to translate evaluation findings into binding launch decisions and embed requirements into development lifecycles.
- Lead applied AI safety and alignment research informing model evaluation, red teaming, and mitigation, while building a network of external research partners.

TRUST AND SAFETY AI RESEARCH SCIENTIST *August 2022 – August 2024*
Spotify, San Francisco, CA

- Led evaluations of LLMs and multimodal models, testing whether policy fine-tuning reliably governed model behavior and identifying where it failed.
- Benchmarked commercial and internal models using public and purpose-built evaluation sets, developing subgroup evaluation methods grounded in social science.
- Translated research on model capabilities and limitations into policy and secured cross-functional agreement across engineering, policy, and trust and safety.
- Founded a standing research series, setting the agenda and recruiting external academic and industry speakers on emerging AI risk topics for engineering, product, and policy audiences.

AI VISITING RESEARCH SCIENTIST *May 2019 – April 2022*
Facebook, Menlo Park, CA

- Established normative positions on generative AI personas, assessing the conditions under which synthetic AI-generated entities should be permitted.
- Conducted applied normative research on operationalizing fairness in production ML systems, collaborating with philosophers and policy researchers to translate contested ethical concepts into implementable system criteria.
- Led evaluations of AI content moderation systems to identify and measure societal biases and implement mitigation strategies.

AI RESEARCH SCIENTIST INTERN *September 2018 – May 2019*
Twitter, San Francisco, CA

- Evaluated hate speech classifiers for subgroup bias and developed thresholding criteria to translate fairness findings into model decisions.
- Adapted academic survey methods to map AI use across Twitter, establishing an early model inventory used to support subsequent model evaluation and risk response.
- Built external research relationships by organizing a FAccT 2019 networking event connecting academic and industry AI researchers.

AI ETHICS RESEARCHER, FIELD GUIDE TO AI FUTURES *December 2018 – May 2019*
Dovetail Labs, Oakland, CA

- Researched global AI development and deployment for Dovetail Labs' Field Guide to AI Futures collaboration with Microsoft.
- Produced an internal research brief synthesizing critical scholarship on the social and political implications of emerging AI systems.
- Authored blog posts examining AI deployment and misuse in high-stakes state contexts, from unvalidated deception detection and AI-enabled border enforcement to dual-use surveillance technologies.

PRIVACY AND TECHNOLOGY RESEARCH FELLOWSHIP *May 2018 – August 2018*
Georgetown Law, Washington, D.C.

- Anticipated civil rights and security risks of emerging AI-enabled surveillance technologies and developed recommendations to inform mitigation efforts.
- Supported programming for Georgetown Law's Color of Surveillance conference and served as a panel moderator.
- Built a cross-sector network of academic, civil society, industry, and government leaders in AI and technology policy.

FELLOWSHIPS

TRANSATLANTIC DIGITAL DEBATES FELLOWSHIP *Fall 2022*
New America's Open Technology Institute & Global Public Policy Institute, Berlin, Germany

- Brought together 18 German and American young professionals from government, business, civil society, academia, and the media for month-long sessions of problem-oriented dialogue on online platform regulation, data protection, disinformation, targeted digital advertising, antitrust, data governance, and innovation policy.

COLLECTIVE SAFETY FELLOWSHIP *July 2018 – May 2019*
Yerba Buena Center for the Arts, San Francisco, California

- Selected to participate in a yearlong process of inquiry, dialog, and project generation around the question "how might we pursue and sustain collective safety?" with a cohort of creative thinkers and artists in the Bay Area.

DIPLOMACY AND DIVERSITY FELLOWSHIP *Summer 2016*
Humanity in Action, Washington D.C., Berlin, Warsaw, Athens

- Transatlantic educational program offering fellows the opportunity to study how American and European governments and societies are responding to international issues including immigration, xenophobia, the EU and the Brexit crisis, security and human rights, and technological change.

AWARDS

TECH FOR SOCIAL GOOD FELLOWSHIP *Spring 2018 – present*

- Awarded a grant to design a project on the biases of facial recognition and advise two undergraduate students to carry out the project.

EUGENE COTA ROBLES FELLOWSHIP FOR GRADUATE STUDY *Fall 2015 – Summer 2019*

- Awarded by UC Berkeley to its top admitted doctoral students to support graduate study for four years in recognition of distinguished academic record.

PUBLICATIONS

Barreto, R. 2026. **Alignment Inertia: Auditing the Durability of Training Data Influence Through Policy Override Resistance**. Under review, NeurIPS Workshop on Attributing Model Behavior at Scale (ATTRIB).

Keyes, O. and Barreto, R. 2026. **Making Race from Face**. In review, Encoding Realities anthology (Palgrave).

Ovalle, E. and Barreto, R. 2026. **Breaking RLAIF**. In progress.

Ashar, Ginena, Barreto, and Cramer. 2024. **Algorithmic Impact Assessments at Scale: Practitioners' Challenges and Needs**. Journal of Online Trust and Safety. Vol 2. No 4.

Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, Chris Kennedy. 2022. **Assessing Annotator Identity Bias via Item Response Theory: A Case Study in a Hate Speech Corpus**. FAccT Proceedings.

Bakalar, Barreto, et al. 2021. **Fairness On The Ground: Applying Algorithmic Fairness Approaches To Production Systems**. Meta AI.

TALKS, WORKSHOPS, AND PUBLIC-FACING WORK

Leveraging Rasch Measurement Theory for Data Perspectivism *June 2022*
Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, Chris Kennedy. 1st Workshop on Perspectivism Approaches to NLP. ACL 2022.

Responsible AI Panel *September 2021*
Panelist. Responsible Tech Summit, All Tech Is Human. Virtual / Remote.

ML and the Construction of Race in Genome Wide Association Studies *May 2021*
Genealogies of Data, Junior Scholar Symposium (Google). Virtual / Remote.

NYU Public Interest Technology Workshop *June 2021*
Virtual / Remote.

AI at the Borderlands *December 2020*
NeurIPS Conference, Resistance AI. Virtual / Remote.

The Operationalization of SESTA-FOSTA Through Machine Learning Models *January 2020*
Doctoral Colloquium. ACM FAT*. Barcelona, Spain.

What does Queer Theory Teach Us about Algorithmic Decisionmaking? *August 2019*
American Sociological Association. New York, NY.

Workshop on Technology, Law, and Society *June 2018*
UC Irvine, CA.

A Shared Lexicon for Research and Practice in Human-Centered Software Systems *May 2018*
Privacy Law Scholars Conference. Washington, D.C.

Workshop on AI and Communication Governance *March 2018*
Alexander von Humboldt Institute for the Internet and Society. Berlin, Germany.

A Shared Lexicon for Research and Practice in Human-Centered Software Systems *February 2018*
Tutorial. ACM FAT* Conference. New York, NY.

Emerging Algorithms, Borders, and Belonging *May 2017*
Essay. Humanity in Action Diplomacy and Diversity Fellowship collection.

PROGRAMMING Proficient in Python, SQL, Bash, scikit-learn, PyTorch