Anaelia Ovalle

Website | anaeliaovalle | anaeliaovalle | anaeliaovalle@gmail.com

Research Interests

Trustworthy ML, Alignment, AI Safety, Evaluation, Social Impacts, Scientific & Clinical Grounding

Education

University of California, Los Angeles Dec 2024
PhD in Computer Science, Advised by Professor Kai-Wei Chang
M.S. Computer Science Sept 2021

University of San Francisco May 2017
B.S. Data Science

Work Experience

Meta
Fundamental AI Research (FAIR), *AI Research Scientist* Dec 2024 - Present

– Spearheaded research on faithful reasoning, developing an evaluation framework for reasoning trace quality that was adopted by other researchers and influenced internal approaches to assessing disparities in language models and vision-language models (to be presented at EMNLP MRL Workshop 2025).

Global Partnership on AI
Equality and Diversity in Artificial Intelligence, *Machine Learning Consultant* Feb 2023 - June 2024

– Provided expert analysis of responsible AI policies to international AI policy leaders, where I identified gaps in articulation of bias detection and mitigation across the AI lifecycle. Delivered actionable recommendations for embedding substantive equality into machine learning systems, with contributions featured in the 2025 OECD AI Policy Report. [URL]

– Partnership with Mila & The Organisation for Economic Co-operation and Development (OECD)

Meta
Fundamental AI Research (FAIR), *AI Research Scientist Intern* Aug 2023 - Dec 2023

– Created new approach leveraging implicit reward signals to systematically measure how post-training techniques (e.g., Direct Preference Optimization) can amplify pre-existing biases in large language models. Published in the 2025 ACM FAccT Conference.

Responsible AI (RAI), *Research Scientist Intern* June 2021 - Oct 2021

– Built a fairness measurement framework for Instagram Reels ranking, ensuring equal opportunity in content sourcing while preserving user relevance. The algorithm was later adopted by the Facebook Ads team.

Amazon

Trustworthy Alexa, *Applied Scientist Intern* June 2023-Aug 2023, June 2022 - Sept 2022

– Discovered how LLM gender biases are exacerbated by BPE Tokenization. Accepted to 2024 NAACL and 2024 Neurips Queer in AI Workshop (top 15%).

– Created a benchmark and automatic metric to quantify gender-diverse bias in LLMs. Published in 2023 ACM FAccT Conference.

– Led dialogue across teams outside of Alexa for stakeholder buy-in and interdisciplinary feedback on experiment design.

Prime Video Compliance, *Applied Scientist Intern* June 2020 - Sept 2020

– Enabled racist content discovery within Amazon Prime Video for improved regional Prime Video compliance and more transparent content flagging for viewer discretion.

Unity Technologies

Monetization Team, *Data Scientist* Sept 2017 - Aug 2019

– Developed deep learning and reinforcement learning algorithms to optimize lifetime value of users in mobile games, resulting in 5% gain in ad spend.

– Partnered with engineering to deploy A/B tests, delivering data-driven recommendations to product managers to optimize in-app purchase and ad placement strategy.

Analytics Team, *Data Scientist Intern* May 2017 - Aug 2017

– Leveraged Unity gameplay data to predict player engagement and inform churn prevention strategies. Implemented ML algorithms such as logistic regression, XGBoost, and neural networks with PySpark and TensorFlow.

Selected Works

[1] Anaelia Ovalle, Candace Ross, Sebastian Ruder, et al. "Beg to Differ: Understanding Reasoning-Answer Misalignment Across Languages". In: *EMNLP Workshop on Multilingual Representation Learning*. To appear. 2025.

[2] Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, et al. "The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models". In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2024. URL: https://dl.acm.org/doi/10.1145/3715275.3732196.

[3] Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, et al. "Tokenization Matters: Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies". In: *Findings of the Association for Computational Linguistics (NAACL Findings)*. Also appeared at NeurIPS 2023 Workshop on Queer in AI; Top 15% of papers. 2024. URL: https://aclanthology.org/2024.findings-naacl.113/.

[4] Organizers of Queer in AI, Anaelia Ovalle, Arjun Subramonian, et al. "Queer in AI: A Case Study in Community-Led Participatory AI". In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Best Paper Award. 2023, pp. 1882–1895. URL: https://dl.acm.org/doi/10.1145/3593013.3594134.

[5] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, et al. ""I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation". In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2023, pp. 1246–1266. URL: https://dl.acm.org/doi/10.1145/3593013.3594078.

[6] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, et al. "Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness". In: *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 2023, pp. 496–511. URL: https://dl.acm.org/doi/10.1145/3600211.3604705.

[7] Nathan Dennler, Anaelia Ovalle, Ashwin Singh, et al. "Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms". In: *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 2023, pp. 375–386. URL: https://dl.acm.org/doi/10.1145/3600211.3604682.

[8] Anaelia Ovalle, Sunipa Dev, Jieyu Zhao, et al. "Auditing Algorithmic Fairness in Machine Learning for Health with Severity-Based LOGAN". In: *Health Intelligence: AI in Health and Medicine*. Ed. by Yu Jin, Chandan Reddy, Joyce C. Ho, et al. Vol. 13942. Cham: Springer, 2023, pp. 146–160. URL: https://doi.org/10.1007/978-3-031-36938-4_10.

[9] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, et al. "Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies". In: *Proceedings of EMNLP*. 2021, pp. 1968–1994. URL: https://aclanthology.org/2021.emnlp-main.150.

[10] Chandra L. Ford, Bita Amani, Nina T. Harawa, et al. "Adequacy of Existing Surveillance Systems to Monitor Racism, Social Stigma and COVID Inequities: A Detailed Assessment and Recommendations". In: *International Journal of Environmental Research and Public Health* 18.24 (2021), p. 13099. URL: https://doi.org/10.3390/ijerph182413099.

Trainings

Intersectionality Training Summer Intensive, Intersectionality Training Institute Aug 2022
Leadership Training, Unity Technologies Aug 2019

Invited Talks

Deep Learning IndabaX Equatorial Guinea Jan 2025
Title: From Algorithms to Impact: How AI Reflects and Shapes our Society

Hub de América Latina y el Caribe, Investigación sobre Inteligencia Artificial Feb 2024
Title: Soy Quien Soy: Queer Voices for Inclusive Language Technologies

Meta, FAIR Society & AI Team March 2023
Title: Intersectionality as a Means for Interrogating Power Structures in AI

UC Berkeley, Data Science for Social Justice Workshop July 2023
Title: Towards Inclusive Harm Evaluation Frameworks: The Case of Gender Minorities

University of Michigan, Gender Diversity in Robotics Seminar July 2023
Title: Queer Realities Meet Harm Evaluation Frameworks for Generative AI

Service

Invited Reviewer, *Cell Patterns* 2025
Program Committee Member, ACM FAccT 2025
Co-Organizer, NAACL Student Research Workshop 2024
Program Committee Member, EACL 2023
Ethics Reviewer, NeurIPS 2023
Co-Organizer, KDD Equity, Diversity & Inclusion Day 2023
ACL Trustworthy NLP Workshop Co-Organizer and Reviewer 2023, 2024

Community Building

Queer in AI Core Organizer Present
Deep Learning Indaba Mentorship Program Committee Present
Deep Learning Indaba Mentor Present
ACM FAccT CRAFT Workshop Organizer 2025, 2022

Honors and Awards

2023 ACM FAccT Best Paper Award 2023
Ford Foundation Research Grant, Queer in AI, $50k 2023
NSF NRT Program Trainee, $22k 2021
Ford Fellowship Honorable Mention 2021
GEM Full Fellowship, $16k, 1 year 2020
Eugene V. Cota Robles Full Fellowship, $30k, 4 years 2019
The Anita Borg Systers Pass It On Award, $3k 2018
USF Magna Cum Laude 2017
USF Senior Leadership Award 2017
USF Changes the World From Here Academic Research Award 2017