Dora Zhao

Work Email: dorothyz@stanford.edu
Permanent Email: dorazhao99@gmail.com
URL: https://dorazhao99.github.io/

Education

2023-Present PhD in Computer Science, Stanford University, GPA: 4.0/4.0
Advisors: Michael S. Bernstein and Diyi Yang

2021-2022 MSE in Computer Science, Princeton University, GPA: 4.0/4.0
Advisors: Olga Russakovsky and Andrés Monroy-Hernández

2017-2021 AB in Computer Science (*summa cum laude*), Princeton University, GPA: 3.97/4.0
Advisor: Olga Russakovsky

Publications

** denotes equal contribution*

Conference papers

[C15] Ali, D., **Zhao, D.**, Koenecke, A., Papakyriakopoulos, O. "Operationalizing Pluralistic Values in Large Language Model Alignment Reveals Trade-offs in Safety, Inclusivity, and Model Behavior" *AAAI Conference on Artificial Intelligence (AI Alignment Track)*, 2026.

[C14] Yang, Y., Lee, C., Feng, S., **Zhao, D.**, Wen, B., Liu, A., Tsvetkov, Y., Howe, B. "Escaping the SpuriVerse: Can Large Vision-Language Models Generalize Beyond Seen Spurious Correlations?" *NeurIPS Datasets and Benchmarks (NeurIPS D&B)*, 2025.

[C13] Kolluri, A., Su, R., Jahanbakhsh, F., **Zhao, D.**, Piccardi, T., Bernstein, M.S. "Alexandria: A Library of Pluralistic Values for Realtime Re-Ranking of Social Media Feeds." *International AAAI Conference on Web and Social Media (ICWSM)*, 2025.

[C12] **Zhao, D.**, Yang, D., Bernstein, M.S. "Knoll: Creating a Knowledge Ecosystem for Large Language Models." *ACM Symposium on User Interface Software and Technology (UIST)*, 2025.

[C11] **Zhao, D.***, Ma, Q.*, Zhao, X., Si, C., Yang, C., Louie, R., Reiter, E., Yang, D., Wu, S. "SPHERE: An Evaluation Card for Human-AI Systems." *Annual Meeting of the Association for Computational Linguistics (ACL Findings)*, 2025.

[C10] **Zhao, D.***, Scheuerman, M.*, Chitre, P., Andrews, J., Panagiotidou, G., Walker, S., Pine, K., Xiang, A. "A Taxonomy of Challenges to Curating Fair Datasets." *NeurIPS Datasets and Benchmarks (NeurIPS D&B) (Oral)*, 2024.
Top 0.5% of papers

[C9] Hirota, Y., Andrews, J., **Zhao, D.**, Papakyriakopoulos, O., Modas, A., Nakashima, Y., Xiang, A. "Resampled Datasets Are Not Enough: Mitigating Societal Bias Beyond Single Attributes." *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[C8] **Zhao, D.**, Andrews, J., Papakyriakopoulos, O., Xiang, A. "Position: Measure Dataset Diversity, Don't Just Claim It." *International Conference on Machine Learning (ICML)*, 2024.
Best Paper Award

[C7] Andrews, J., **Zhao, D.***, Thong, W.*, Modas, A.*, Papakyriakopoulos, O.*, Nagpal, S.*, Xiang, A. "Ethical Considerations for Responsible Data Curation." *Neural Information Processing Systems Datasets and Benchmarks (NeurIPS D&B) (Oral)*, 2023.
Top 0.5% of papers

[C6] Ramaswamy, V.V., Lin, P., **Zhao, D.**, Adcock, A., van der Maaten, V., Ghadiyaram, D., Russakovsky, O. "GeoDE: a Geographically Diverse Evaluation Dataset for Object Recognition." *Neural Information Processing Systems Datasets and Benchmarks (NeurIPS D&B)*, 2023.

[C5] **Zhao, D.***, Meister, N.*, Wang. A, Ramaswamy, V.V., Fong, R., and Russakovsky, O. "Gender Artifacts in Visual Datasets." *International Conference on Computer Vision (ICCV)*, 2023.

[C4] **Zhao, D.**, Andrews, J., Xiang, A. "Men Also Do Laundry: Multi-Attribute Bias Amplification." *International Conference on Machine Learning (ICML)*, 2023.

[C3] Papakyriakopoulos, O., Choi, A., Thong, W., **Zhao, D.**, Andrews, J., Bourke, R., Xiang, A., Koenecke, A. "Augmented Datasheets for Speech Datasets and Ethical Decision-Making." *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[C2] **Zhao, D.**, Inaba M., and Monroy-Hernández, A. "Understanding Teenage Perceptions and Configurations of Privacy on Instagram." *Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 2022.

[C1] **Zhao, D.**, Wang. A, and Russakovsky, O. "Understanding and Evaluating Racial Biases in Image Captioning." *International Conference on Computer Vision (ICCV)*, 2021.

Journal articles

[J2] Xiang, A., Andrews, J. T. A., Bourke, R. L., Thong, W., LaChance, J. M., Georgievski, T., Modas, A., Rahmattalabbi, A., Ba, Y., Nagpal, S., Papakyriakopoulos, O., **Zhao, D.**, Xue, J., Matthews, V., Gong, L., Hoag, A. T., Cimpoi, M., Sankaranarayanan, S., Hutiri, W., Scheuerman, M. K., Abedi, A. S., Stone, P., Wurman, P. R., Kitano, H., Spranger, M. "Fair human-centric image dataset for ethical AI benchmarking" *Nature*, 2025.

[J1] Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., **Zhao, D.**, Shirai, I., Narayanan, A., and Russakovsky, O. "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets." *International Journal of Computer Vision (IJCV)*, 2022.

Workshop papers

[W2] Ali, D., Kocak, A., **Zhao, D.**, Koenecke, A., Papakyriakopoulos, O. "A Sociotechnical Perspective on Aligning AI with Pluralistic Human Values." *ICLR Workshop on Bidirectional Human-AI Alignment*, 2025.
Best Poster Award

[W1] One of nineteen authors*. "Anti-Racist HCI: Notes on an Emerging Critical Technical Practice" *alt.chi,* 2022.

Under Submission

[S4] **Zhao, D.**, Cha, H., Ryan, M.J., Wang, A., Baker-Ramos, R., Helekahi-Kaiwi, E., Diego, R., Yang, D. "Whose Knowledge Counts? Co-Designing Community-Centered Auditing Tools with Educators in Hawai'i." 2025.

[S3] Zhang Y., **Zhao, D.**, Hancock, J., Kraut, R., Yang, D. "The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being." 2025.

[S2] Scheuerman, M., **Zhao, D.**, Andrews, J., Panagiotidou, G., Chitre, P., Walker, S., Pine, K., Xiang, A. "Fairness in Context: A Conceptual Framework for Guiding Fairness Decisions in Machine Learning Dataset Curation." 2025.

[S1] **Zhao, D.***, Jahanbakhsh, F.*, Piccardi, T., Robertson, Z., Epstein, Z., Koyejo, S., Bernstein, M.S. "Value Alignment of Social Media Ranking Algorithms." 2025.

Preprints

[P1] **Zhao, D.**, Yang, D., Bernstein, M.S. "Mapping the Spiral of Silence: Surveying Unspoken Opinions in Online Communities" 2024.

Presentations and Talks

A Taxonomy of Challenges to Curating Fair Datasets

- Invited talk at University of British Columbia's Statistics Equity, Diversity and Inclusion Speaker Series
- Oral talk at *NeurIPS 2024* – New Data & Safety Session
- UT Austin, Responsible Data Management Course (Instr. Hanlin Li)

Encoding Basic Human Values in Social Media Feed Ranking Algorithms

- Invited talk at Brown University's CNTR Seminar on Technology and Society

Measure Dataset Diversity, Don't Just Claim It

- Invited talk at Sony Research's AI Ethics Reading Group
- Oral talk at *ICML 2024* — Data & Society Session

Multimodal Systems through a Social Lens: Uncovering and Mitigating Biases

- Invited talk at *CIKM 2023 Multimodal Human Understanding for the Web and Social Media Workshop*

Understanding Teenage Perceptions and Configurations of Privacy on Instagram

- Part of a panel at *CSCW 2022 Northeast Meetup*

Understanding and Evaluating Racial Biases in Image Captioning

- Part of an invited talk at *CVPR 2022 AI for Content Creation (AI4CC) Workshop*
- Part of an invited talk at *CVPR 2021 Visual Question Answering (VQA) Workshop*

Work Experience

July 2022 - Aug 2023 **AI Engineer**, *Sony Research (AI Ethics Team)*, New York, NY

- Assisted with large-scale data collection for an ethical human-centric dataset by creating toolkits for data quality review
- Developed educational materials on fairness, bias, and privacy concerns in AI made publicly available for users of Sony Semiconductor Solutions Corporation's edge AI sensing platform, AITRIOS
- Created AI Ethics Checklists and designed a research ethics review process that has been adopted by all flagship projects in Sony AI

May 2021 - Aug 2021 **Research Intern**, *Sony Research (AI Ethics Team)*, Remote

- Proposed methods for using generative adversarial networks to reduce spurious correlations in image datasets
- Mentors: Alice Xiang, Jerone Andrews

Sept 2017 - Dec 2018 **Undergraduate Research Assistant**, *Stigma and Social Perception Laboratory*, Princeton, NJ

- Assisted Professors Nicole Shelton and Stacey Sinclair on audit study exploring the representations of diversity in college admissions materials and mission statements

Teaching

Spring 2022 **Advanced Programming Techniques (COS 333)**, *Teaching Assistant*, Princeton University
Fall 2021 **Advanced Programming Techniques (COS 333)**, *Teaching Assistant*, Princeton University
Spring 2021 **Computer Vision for Social Good (COS IW 07)**, *Teaching Assistant*, Princeton University
Spring 2021 **Fairness in Visual Recognition (COS IW 08)**, *Teaching Assistant*, Princeton University

Honors and Awards

2025 **Paul and Daisy Soros Fellowship for New Americans**

2024 **Best Paper at ICML 2024**
For "Position: Measure Dataset Diversity, Don't Just Claim It"

2024 **Brown Institute for Media Innovation Magic Grant**
Awarded $60k for research on the spiral of silence on social media

2021 **Joseph Clifton Elgin Prize**
Award for the senior in Princeton University's School of Engineering and Applied Sciences who has done the most to advance the interests of the School in the community at large

2021 **Sigma Xi Book Award for Outstanding Undergraduate Research**

2020 **Computing Research Association's Outstanding Undergraduate Researcher – Honorable Mention**

2020 **Accenture Prize in Computer Science**
Merit-based award for top 2 seniors in the Princeton Computer Science department

Professional Service

Organizing Committee

CSCW 2025 – Workshop on Responsibly Training Foundation Models

Program Committee & Reviewing

Conferences
UIST (2025), CHI (2025*, 2026), CSCW (2024*, 2025), Web (2024), CVPR (2023), AAAI (2022)
** recognition for outstanding review(s)*

Workshops
CVPR 2024 Workshop on Responsible Data
ECCV 2024 Critical Evaluation of Generative Models and their Impact on Society

Outreach

2024-2025 **Stanford Human-Computer Interaction Seminar**, *Organizer* - Help organize weekly seminar series on topics related to HCI

2023-2024 **Stanford Computing and Society Group**, *Organizer* - Host biweekly speaker seminars on topics related to computing and society

2023-2025 **Stanford CS Undergraduate Mentoring Program**, *Volunteer* - Mentor Stanford CS undergraduate students from underrepresented backgrounds

2023-Present **Stanford CS Student-Applicant Support Program**, *Volunteer* - Provide feedback to first-generation, low-income students on their application materials before they formally apply to the Stanford CS PhD program

Mentoring

2025-Present Hannah Cha, *Stanford CS MS*, project on auditing biases in LLMs for Hawaiian classrooms [S4]
2024-2025 Renn Su, *Stanford CS Undergrad*, project on ranking social media feed based on values [C12]
2024-2025 Akaash Kolluri, *Stanford CS Undergrad*, project on ranking social media feed based on values [C12]
2024-Present Yutong Zhang, *Stanford CS MS*, project on social implications of AI companions [S3]

Last updated: November 8, 2025