Robert Wolfe

wolferobert3@gmail.com
wolferobert3.github.io
www.github.com/wolferobert3

Education

Ph.D., Information Science, in progress
University of Washington – Seattle, WA
GPA: 3.93
Coursework: Quantitative Methods, Qualitative Methods, Research Design, Design Inquiry, Theory of Information, Problematic Information, Critical and Cultural Perspectives in Information Science

M.S., Computer Science, 2021
The George Washington University – Washington, D.C.
Thesis: VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models, published at AAAI 2022
GPA: 4.0
Coursework: Machine Learning; Statistical NLP; Advanced Algorithms; Computer Architecture; Programming Languages; Computer Systems

M.A., English Literature, 2014
Georgetown University – Washington, D.C.
GPA: 3.9
Thesis: Driven by Difference (a study of rhetorical interchange among state-level U.S. secession groups)
Fellow at the Center for New Designs in Learning and Scholarship

B.A., English Literature, 2012
University of Maryland, College Park
GPA: 3.78
Phi Beta Kappa, University Honors, Jimenez-Porter Writer’s House Citation
Graduate of the Undergraduate Technology Apprenticeship Program (UTAP)

Work Experience

Deep Learning Research Intern, AKASA, San Francisco, CA (Remote)
January-March 2023

- Developed a generative language model for radiologists using the imaging notes included in the MIMIC dataset. Prototyped models trained from scratch, fine-tuned from pretrained base, and adapted using low-rank adaptation methods. Developed custom tokenization for domain.
- Developed a novel loss for language modeling based on the Barlow Twins algorithm used for self-supervised representation learning in computer vision.
- Supported engineering team in project to predict medical codes from physicians’ notes using long-sequence language models, such as longT5.

DisputeSoft: Software Dispute Experts, Washington, D.C. Office
2014-2021

- Director of marketing, operations, and HR.
- Authored expert witness memos on complex legal consulting matters.

Teaching

Note: all teaching experience is at the University of Washington iSchool.

2024 – Spring

IMT598: Epistemological Foundations of AI. With Dr. Bill Howe. Gave invited discussion on **recent approaches to the use of generative AI in professional fact-checking**, contextualizing uses within class materials on AI epistemologies.

2023 – Spring

INFO371: Advanced Methods in Data Science. Teaching Assistant to Dr. Ott Toomet. Managing lab sessions, updating programming and data analysis assignments, grading. Course content covers causality, machine learning, introductory Bayesian statistics, and introductory deep learning, including the fundamentals of computer vision and NLP**.** Spring 2023. **Student Rating: 4.9/5 (top 10-20% of TAs during Spring 2023 at the University of Washington).**

2023 – Winter

INFO466: Moral Reasoning and Interaction Design. With Dr. Alexis Hiniker. Developing lecture and class materials to support students in **learning about the ethical consequences of technical design decisions**, with particular attention to generative AI. Gave a lecture on pragmatism in design and ran a class workshop. Winter 2023.

INFO370: Core Methods in Data Science. Teaching Assistant to Dr. Ott Toomet. Managing lab sessions, updating programming and data analysis assignments, grading. Course content covers data analysis, statistical inference, regression, and basic machine learning. Winter 2023. **Student Rating: 4.9/5 (top 10-20% of TAs during Winter 2023 at the University of Washington).**

2022 – Fall

INFO270: Data Reasoning. Teaching Assistant to Dr. Jevin West and Dr. Carl Bergstrom. **Developed and gave two class lectures** on AI bias and overhype, **created Jupyter notebook** used as course material **for teaching statistical bias** in AI to all discussion sections, participated in **panel discussion on scientific uses of AI** with course professors, taught discussion sessions, graded assignments. Fall 2022. **Student Rating: 4.6/5.**

2022 – Spring

IMT575: Machine Learning 3: Applications, Scaling, and Ethics. With Dr. Aylin Caliskan. Developed and gave the **invited guest lecture on Data Science for Social Good,** covering historical trends on applications of AI for good, and recent research on bias in AI. March 2022.

Publications

Below follows my list of publications.

2024

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, and Bill Howe. Laboratory-Scale AI: Open-Weight Models are Competitive Even in Low-Resource Settings. ACM Conference on Fairness, Accountability, and Transparency (FAccT) 2024.
Empirical study of the viability of small, open generative models as an alternative to large, proprietary, closed models. Finds that open models are competitive given a relatively small amount of data, and offer benefits in the form of cost-efficiency, differential privacy, and tunable abstention properties (reducing hallucination). **23% acceptance rate.**

Robert Wolfe and Tanushree Mitra. The Impact and Opportunities of Generative AI in Fact-Checking. ACM FAccT 2024.
Interview study with *N*=38 fact-checkers that catalogues the in-use, in-progress, and envisioned uses of generative AI in fact-checking, along with epistemic challenges preventing further use. Introduces the dimension of Verification to the design space of generative AI, and takes a value-sensitive approach to mapping tensions between generative AI and fact-checking. **23% acceptance rate.**

Robert Wolfe and Alexis Hiniker. Expertise Fog on the GPT Store: Deceptive Design Patterns in User-Facing Generative AI. ACM CHI 2024 Workshop on Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices.
Position paper with light empirical results arguing that the design of the OpenAI GPT Store encourages deceptive design patterns related to the presentation of expertise in customized versions of ChatGPT.

Aayushi Dangol, Michelle Newman, Robert Wolfe, Jin Ha Lee, Jason Yip, Julie Kientz, and Caroline Pitt. Mediating Culture: Cultivating Socio-cultural Understanding of AI in Children through Participatory Design. ACM Conference on Designing Interactive Systems (DIS) 2024.
Introduces participatory approach to co-designing AI with kids in ways that facilitate an understanding of AI as a mediator of culture. My involvement included building a prototype vision-language AI system and helping to run human subjects sessions.

Yiwei Yang, Anthony Zhe Liu, Robert Wolfe, Aylin Caliskan, and Bill Howe. Label-Efficient Group Robustness via Out-of-Distribution Concept Curation. Computer Vision and Pattern Recognition (CVPR) 2024.
Introduces a Concept Distributively Robust Optimization (DRO) framework that takes curated sets of images for a given concept to estimate group labels, and uses those labels to train with a SOTA DRO objective, significantly reducing classifier biases with relatively small, manually curated sets of images.

2023 **Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. ACM FAccT 2023.**
Mixed methods study of sexually objectifying biases that attend representations of women and girls in multimodal AI models. Traces biases from the embedding space of 9 CLIP models, and in the output of generative text-to-image models such as VQGAN-CLIP and Stable Diffusion. **25% acceptance rate.**

Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. Evaluating Biased Attitude Associations of Language Models in an Intersectional Context. AI Ethics and Society (AIES) 2023.
Introduces an SVM-based method for learning subspaces corresponding to human attitudes and concepts in causal and bidirectional transformer language models, and analyzes the human attitudes and biases reflected in the models in an intersectional context. Expanded the scope of the experiments to four additional language models, substantially revised paper, and contributed research code and data.

Yiwei Yang, Anthony Zhe Liu, Robert Wolfe, Aylin Caliskan, and Bill Howe. Regularizing Model Gradients with Concepts to Improve Robustness to Spurious Correlations. ICML Workshop on Spurious Correlations, Invariance, and Stability. 2023.
Proposes a method known as CReg to penalize a machine learning model's sensitivity to a protected attribute, including the absence of group labels at the dataset level, outperforming the use of Empirical Risk Minimization (ERM) for regularization. Contributed methodological and conceptual ideas.

2022 **Robert Wolfe and Aylin Caliskan. American==White in Multimodal Language-and-Image AI. AIES 2022.**
Study of biases in three multimodal language-and-image AI models: CLIP, SLIP, and BLIP. Shows that language-and-image AI learns statistically veridical information about state-level demographic distributions. Also demonstrates that some regions, such as the U.S., become associated in AI with a dominant social group – in this case, White individuals. **Selected to open the conference, top 1% of peer reviews.**

Robert Wolfe and Aylin Caliskan. Contrastive Multimodal Pretraining Magnifies the Semantics of Natural Language Representations. Association for Computational Linguistics (ACL) 2022.
Intrinsic evaluation of the surprising properties of contextualized word embeddings and sentence embeddings formed by the CLIP text encoder in comparison with those formed by GPT-2. **Selected as an oral presentation** (**top ~8% of accepted papers**).

Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. Evidence for Hypodescent in Visual Semantic AI. ACM FAccT 2022.

Study of biases in CLIP as they pertain to the perception of multiracial individuals. Used a generative adversarial network to replicate an experiment from experimental psychology, and showed that CLIP has learned an analogue of the rule of hypodescent, or one-drop rule. **19% acceptance rate for track submitted**, 26% overall. **Selected for inclusion in KUDOS platform**.

Robert Wolfe and Aylin Caliskan. Markedness in Visual Semantic AI. ACM FAccT 2022.
Study of biases in the multimodal language-and-image AI model CLIP. Examines the proclivity of CLIP to mark the race, gender, and age of some individuals while leaving it unmarked for dominant social groups. **23% acceptance rate for track submitted**, 26% overall.

Robert Wolfe and Aylin Caliskan. VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models. Association for the Advancement of Artificial Intelligence (AAAI) 2022.
Method for intrinsic evaluation of contextualized word embeddings using human-rated psycholinguistic measurements. Originally a master's thesis at GWU. **15% acceptance rate.**

Robert Wolfe and Aylin Caliskan. Detecting Emerging Associations and Behaviors Using Regional and Diachronic Word Embeddings. IEEE International Conference on Semantic Computing (ICSC) 2022.
Methods based on the word embedding association test (WEAT) for detecting changes in semantics over time in Dirichlet-smoothed word embeddings trained on low-resource Twitter corpora.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. Gender Bias in Word Embeddings: A Comprehensive Overview of Syntax, Frequency, and Semantics. AIES. 2022.
Study of gender biases in GloVe and fastText word embeddings taking into account the many properties of language often overlooked when examining word embedding bias. Wrote the first draft of the paper, contributed code and data for the project, mentored junior scientist.

2021

Robert Wolfe and Aylin Caliskan. Low-Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. Empirical Methods in Natural Language Processing (EMNLP) 2021.
Study of bias in contextualized word embeddings based on the correspondence of frequency, intrinsic bias, and self-similarity. **26% acceptance rate**.

Invited Talks

Robert Wolfe and Aylin Caliskan. *Quantifying Biases and Societal Defaults in Word Embeddings and Language-Vision AI*. **AI Metrology Colloquium Series. National Institute of Standards and Technology.** August 2022.

Robert Wolfe. *Overview of Research in Language-and-Image AI, with Focus on Evidence for Hypodescent in Visual Semantic AI*. **Language as a Window into Human Minds. The Santa Fe Institute for Complex Systems.** June 2022.

Grants and Awards

UW iSchool Strategic Research Initiative Award: Awarded $15,000 grant co-written with PI Dr. Bill Howe, entitled ***Laboratory-Scale AI***. Proposes empirical validation of domain-specific, instruction-tuned open models for their competitiveness with large, general, proprietary models like GPT-4. December 2023. Research published at ACM FAccT 2024.

Google Research: Awarded $60,000 grant co-written with PI Dr. Alexis Hiniker, entitled ***Encouraging Nonviolent Communication in Online Messaging Platforms***. Proposes user-centered design of AI-driven technologies for promoting empathy and nonviolence. April 2023. Research under submission.

Press Coverage

Business Insider, 2023/04/25: *Stable Diffusion and DALL-E display bias when prompted for artwork of 'African workers' versus 'European workers'.* By Thomas Maxwell. Refers to Markedness in Visual Semantic AI.

The Intercept, 2023/04/22: *AI Art Sites Censor Prompts About Abortion*. By Debbie Nathan. Refers to Gender Bias in Word Embeddings.

Insider, 2023/01/16: *ChatGPT could be used for good, but like many other AI models, it's rife with racist and discriminatory bias*. By Hannah Getahun. Refers to Evidence for Hypodescent in Visual Semantic AI.

MIT Tech Review, 2022/12/12: *The viral AI avatar app Lensa undressed me – without my consent*. By Melissa Heikkila. Refers to Markedness in Visual Semantic AI.

Academic Service

Program Committee, AI Ethics and Society, 2024.
Program Committee, AAAI, 2024.
Reviewer, Association for Computational Linguistics Rolling Review, 2024.
Reviewer, NeurIPS, 2024.
Program Committee, AAAI, 2023.
Reviewer, NeurIPS, 2023.
Reviewer, Nature Humanities and Social Science, 2023.
Program Committee, ACM FAccT, 2022.
Reviewer, AI Ethics and Society, 2022.
Reviewer, NeurIPS, 2022.
Reviewer, ICML, 2022.

Community Involvement and Leadership

University Children's Development School, Seattle, WA, 2023. Developed a system for allowing children to interact with multimodal bias and reason about

AI systems and AI fairness in a controlled context. **System adopted as part of the curriculum for upper-elementary school children.**

Member of the Volitional AI Lab, UW iSchool, 2023-present.

Member of the User Empowerment Lab, UW iSchool, 2022-present.

Member of the Implicit Machine Cognition Lab, UW iSchool, 2021-2022.

Responsible AI Systems and Experiences (RAISE). Student Organizing Committee. University of Washington. 2021-2022. Organized RAISE speaker series in coordination with student volunteers and RAISE faculty.

Trail Ranger. Montgomery County, MD. 2018-2021. Assumed primary responsibility for condition of a system of local trails, in coordination with county officials.

Memberships

Association for the Advancement of Artificial Intelligence (AAAI)
Association for Computational Linguistics (ACL)
Association for Computing Machinery (ACM)
Institute of Electrical and Electronics Engineers (IEEE)

Technical Background

Programming: Python, R, C++, Java, C, Julia, PHP, Javascript, CSS/HTML.

Libraries: PyTorch, Tensorflow, Transformers, Diffusers, SK-Learn, Pandas, Keras, NumPy, Gensim, GloVe, fastText, SciPy, StyleGAN, Seaborn, Matplotlib.

Technologies: Git, SQL, AWS, GCP, Bash, Slurm, LaTeX.

Graduate Coursework

Machine Learning ♦ Statistical NLP ♦ Computer Architecture ♦ Computer Systems ♦ Advanced Algorithms ♦ Programming Languages ♦ Object-Oriented Programming ♦ Quantitative Methods ♦ Qualitative Methods ♦ Information Theory ♦ Design Methods ♦ Research Design ♦ Problematic Information ♦ Critical and Cultural Perspectives on Information Science