School of Information
Mary Gates Hall, Ste. 370
University of Washington
Seattle, WA 98195

(301) 331-3163 ☎
rwolfe3[at]uw.edu ✉
wolferobert3.github.io 🌐
github.com/wolferobert3

Robert Wolfe

Education

2021 – 2025 **University of Washington** – Seattle, WA
Ph.D., Information Science – Anticipated Spring 2025
M.S., Information Science
Dissertation: *Approaches to Epistemic Risk in Generative and General Purpose AI*
Committee: Dr. Bill Howe (Advisor), Dr. Alexis Hiniker (Advisor), Dr. Tanushree Mitra, Dr. Leilani Battle (GSR)

2019 – 2021 **The George Washington University** – Washington, D.C.
M.S., Computer Science
Thesis: *The Valence-Assessing Semantics Test for Contextualizing Language Models*
Committee: Dr. Aylin Caliskan (Advisor), Dr. Robert Pless, Dr. Abdou Youssef

2012 – 2014 **Georgetown University** – Washington, D.C.
M.A., English Literature
Thesis: *Driven by Difference: The Embodiment of the Western Maryland Initiative*
Committee: Dr. Patricia O'Connor (Advisor), Dr. Matthew Pavesich

2008 – 2012 **University of Maryland** – College Park, MD
B.A., English Literature
Phi Beta Kappa, University Honors, Technology Apprentice Graduate

Peer-Reviewed Conference Publications

2024 [c.19] **Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings.**
ACM Conference on Fairness, Transparency, and Accountability (FAccT) 2024.
Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, and Bill Howe.
Summary: 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).
Acceptance Rate: 24% Overall, 22% Systems Track.

[c.18] **The Impact and Opportunities of Generative AI in Fact-Checking**
ACM FAccT 2024
Robert Wolfe and Tanushree Mitra.
Summary: 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.
Acceptance Rate: 24% Overall, 28% Users and Experiences Track.

[c.17] **The Implications of Open Generative Models in Human-Centered Data Science Work: A Case Study with Fact-Checking Organizations.**
AI Ethics and Society (AIES) 2024.
Robert Wolfe and Tanushree Mitra.
Summary: Study of open generative models in fact-checking organizations, contextualizing the motivations and challenges of fact-checkers within the data science pipelines used by these organizations.
Acceptance Rate: 31.8% Overall.

[c.16] **Dataset Scale and Societal Consistency Mediate Facial Impression Bias in Vision-Language AI.**
AIES 2024.
Robert Wolfe, Aayushi Dangol, Bill Howe, and Alexis Hiniker.
Summary: Study of the factors affecting the presence of facial impression bias in 43 multimodal CLIP models, as well as the reproduction of facial impression biases by generative multimodal models such as Stable Diffusion.
Acceptance Rate: 31.8% Overall.

[c.15] **Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study.**
AIES 2024.
Robert Wolfe, Aayushi Dangol, Bill Howe, and Alexis Hiniker.
Summary: Study comparing biases about adolescents learned by AI to similar biases identified in traditional and news media sources in both the U.S. and Nepal. Conducts workshops with 13 U.S. teenagers and 18 Nepalese teenagers to understand how teenagers themselves view fair representation in media and AI.
Acceptance Rate: 31.8% Overall.

[c.14] **ML-EAT: A Multilevel Embedding Association Test for Interpretable and Transparent Social Science.**
AIES 2024.
Robert Wolfe, Alexis Hiniker, and Bill Howe.
Summary: Research introducing the Multilevel Embedding Association Test (ML-EAT), a method designed to address issues of ambiguity and difficulty in interpreting the traditional EAT measurement by quantifying bias at several levels of increasing granularity.
Acceptance Rate: 31.8% Overall.

[c.13] **Label-Efficient Group Robustness via Out-of-Distribution Concept Curation.**

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024.

Yiwei Yang, Anthony Zhe Liu, Robert Wolfe, Aylin Caliskan, and Bill Howe.

Summary: 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.

Acceptance Rate: 24% Overall.

[c.12] **Mediating Culture: Cultivating Socio-cultural Understanding of AI in Children through Participatory Design.**

ACM Conference on Designing Interactive Systems (DIS) 2024.

Aayushi Dangol, Michelle Newman, Robert Wolfe, Jin Ha Lee, Jason Yip, Julie Kientz, and Caroline Pitt.

Summary: 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, as well as helping to run human subjects sessions.

Acceptance Rate: 27% Overall.

[c.11] **"Sharing, Not Showing Off": How BeReal Encourages Authentic Self-Presentation on Social Media Through Its Design.**

ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW) 2024.

JaeWon Kim, Robert Wolfe, Ishita Chordia, Katie Davis, and Alexis Hiniker.

Summary: Introduces a set of design guidelines for creating social media platforms that support authentic self-presentation online, such as scaffolding reciprocity and expanding beyond spontaneous photo-sharing to allow users to more accurately and comfortably portray themselves.

2023 [c.10] **Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias.**

ACM FAccT 2023.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan.

Summary: 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.

Acceptance Rate: 25% Overall.

[c.9] **Evaluating Biased Attitude Associations of Language Models in an Intersectional Context.**

AIES 2023.

Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan.

Summary: 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.

Acceptance Rate: 28% Overall.

2022 [c.8] **Contrastive Multimodal Pretraining Magnifies the Semantics of Natural Language Representations.**

Proceedings of the Association for Computational Linguistics (ACL) 2022.

Robert Wolfe and Aylin Caliskan.

Summary: 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.

Acceptance Rate: 21% Overall. Selected for **Oral Presentation** (Top 8% of papers).

[c.7] **VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models.**
Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) 2022.
Robert Wolfe and Aylin Caliskan.
Summary: Method for intrinsic evaluation of contextualized word embeddings using human-rated psycholinguistic measurements. Originally a master's thesis at GWU.
Acceptance Rate: 15% Overall.

[c.6] **Evidence for Hypodescent in Visual Semantic AI.**
ACM FAccT 2022.
Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan.
Summary: 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.
Acceptance Rate: 26% Overall, 19% for Track.

[c.5] **Markedness in Visual Semantic AI.**
ACM FAccT 2022.
Robert Wolfe and Aylin Caliskan.
Summary: 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.
Acceptance Rate: 26% Overall, 23% for Track.

[c.4] **American==White in Multimodal Language-and-Image AI**.
AIES 2022.
Robert Wolfe and Aylin Caliskan.
Summary: 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.
Acceptance Rate: 34%. Selected to **Open the Conference.**

[c.3] **Gender Bias in Word Embeddings: A Comprehensive Overview of Syntax, Frequency, and Semantics.**
AIES 2022.
Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji.
Summary: 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.
Acceptance Rate: 34% Overall.

[c.2] **Detecting Emerging Associations and Behaviors Using Regional and Diachronic Word Embeddings.**
International Conference on Semantic Computing (ICSC) 2022.
Robert Wolfe and Aylin Caliskan.
Summary: 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.

2021 [c.1] **Low-Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models.**
Proceedings of the ACL: Empirical Methods in Natural Language Processing (EMNLP) 2021.
Robert Wolfe and Aylin Caliskan.
Summary: Study of bias in contextualized word embeddings based on the correspondence of frequency, intrinsic bias, and self-similarity.
Acceptance Rate: 26% Overall.

Workshop Publications

2024 [w.2] **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.
Robert Wolfe and Alexis Hiniker.
Summary: 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.

2023 [w.1] **Regularizing Model Gradients with Concepts to Improve Robustness to Spurious Correlations.**
ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability.
Yiwei Yang, Anthony Zhe Liu, Robert Wolfe, Aylin Caliskan, and Bill Howe.
Summary: 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.

Under Submission in Peer-Reviewed Venues

[u.6] **Toward Nonviolent Design: Co-Designing a Human-Centered Framework for AI-Mediated Communication.**
Robert Wolfe, Aayushi Dangol, Jaewon Kim, and Alexis Hiniker.
Summary: Study introducing a human-centered framework for AI-mediated communication that builds on the principles of Nonviolent Communication (NVC).

[u.5] **Volitional AI: The Epistemic Risks of Generative AI Demand Visible Leadership.**
Robert Wolfe, Tanushree Mitra, and Bill Howe.
Summary: Position paper regarding the unique impact of generative AI on the public trust. Argues that generative and general-purpose models must be approached as epistemic infrastructure.

[u.4] **Reading AI and Reading the World: Using an Interactive AI System to Promote Children's Understanding of AI Bias.**
Aayushi Dangol, Robert Wolfe, Akeiylah Dewitt, Ben Chickadel, Julie Kientz, and Sayamindu Dasgupta.
Summary: Introduces the interactive CLIP4KIDS system and studies how students understand AI biases in terms of ”assumptions” and ”stereotypes,” drawing connections between historical injustices and present implicit biases in AI models.

[u.3] **Building the Beloved Community: Designing Technologies for Neighborhood Safety.**
Ishita Chordia, Robert Wolfe, Carl DiSalvo, Jason Yip, and Alexis Hiniker.
Summary: Study probing the development of justice-oriented safety technologies to support collective welfare rather than individualistic notions of safety. Leverages principles of Transformative Justice to offer a novel conceptualization of safety focused on the basic needs of the community.

[u.2] **The Fallacy of 'Likes': Supporting Adolescents in Building Trust with Peers on Social Media.**
JaeWon Kim, Robert Wolfe, Ramya Bhagirathi Subramanian, Mei-Hsuan Lee, and Alexis Hiniker.
Summary: Study probing how social media platforms can support adolescents in building trusting relationships with peers by replacing generic ”Likes” with affordances for precise, effortful reactions, focusing on what matters most to them.

[u.1] **Privacy as Social Norm: Systematically Empowering Teen Privacy Management on Social Media.**
JaeWon Kim, Soobin Cho, Robert Wolfe, Jishnu Nair, and Alexis Hiniker.
Summary: Study identifying sources of fear among teens on social media. Suggests strategies for social media platforms to mitigate dysfunctional fear without compromising privacy.

Invited Research Talks

August 2022 **Quantifying Biases and Societal Defaults in Word Embeddings and Language-Vision AI**.
National Institute of Standards and Technology. AI Metrology Colloquium Series.
Talk and discussion of recent research on biases in vision-language AI models, and comprehensive analyses of gender bias in word embeddings.

June 2022 **Manifestations of Implicit Biases in Language-and-Image AI**.
The Santa Fe Institute for Complex Systems. Language as a Window into Human Minds.
Overview of recent research for a small group of psychologists, decision scientists, and computer scientists at a meeting intended to foster interdisciplinary perspectives.

Teaching Experience

Spring 2023 **Teaching Assistant, Informatics 371: Advanced Methods in Data Science (University of Washington)**.
*Average student rating: 4.9/5. (**Top 10-20% of TAs** at the University of Washington).*
Summary of Role: 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. TA to Dr. Ott Toomet.

Winter 2023 **Teaching Assistant, Informatics 371: Core Methods in Data Science (University of Washington)**.
*Average student rating: 4.9/5. (**Top 10-20% of TAs** at the University of Washington).*
Summary of Role: Managing lab sessions, updating programming and data analysis assignments, grading. Course content covers data analysis, statistical inference, regression, and basic machine learning. TA to Dr. Ott Toomet.

Fall 2022 **Teaching Assistant, Informatics 270: Data Reasoning (University of Washington)**.
Average student rating: 4.6/5.
Developed and gave two class lectures on AI bias and epistemic risk; **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. TA to Dr. Jevin West and Dr. Carl Bergstrom.

Invited Classroom Lectures and Discussions

Spring 2024 **Invited Class Discussion, Information Management 598: Epistemological Foundations of AI (University of Washington)**.
Gave invited discussion on recent approaches to the use of generative AI in professional fact-checking, contextualizing uses within class materials on AI epistemologies. With Dr. Bill Howe.

Winter 2023 **Invited Guest Lecture and Workshop, Informatics 466: Moral Reasoning and Interaction Design (University of Washington)**.
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. With Dr. Alexis Hiniker.

Spring 2022 **Invited Guest Lecture, Information Management 575: Machine Learning 3: Applications, Scaling, and Ethics (University of Washington)**.
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.

Grants, Awards, and Honors

2024 **UW iSchool Conference Travel Award**: $2,000.
Awarded $2,000 to cover conference travel for presentation of two first-author papers at ACM FAccT 2024.

UW iSchool Conference Travel Award: $658.
Awarded $658 to cover expenses related to travel for participation in the ACM CHI 2024 Dark Patterns workshop.

2023 **UW iSchool Strategic Research Initiative Award**: $15,000.
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. **Resulting research published at ACM FAccT 2024.**

Google Research: $60,000.
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. Research under submission.

2022 **UW iSchool Conference Travel Award**: $2,000.
Awarded $15,000 to cover conference travel for presentation of two first-author papers at ACM FAccT 2022.

2012-2014 **CNDLS Fellowship, Georgetown University**: Full tuition fellowship and stipend.
Full tuition fellowship and stipend covering two years in the English M.A. program. Connected to Fellow position at the Georgetown Center for New Designs in Learning and Scholarship (CNDLS), an organization to further technological approaches to higher education.

2008-2010 **Dean's Scholarship, University of Maryland**: $5,000 yearly.
Awarded $5,000 yearly tuition scholarship for the first two years of undergraduate studies.

Press Coverage

April 2023 **Business Insider**. 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 (published at ACM FAccT 2022).

April 2023 **The Intercept**. AI Art Sites Censor Prompts About Abortion. By Debbie Nathan.
Refers to Gender Bias in Word Embeddings (published at AIES 2022).

January 2023 **Insider**. 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 Markedness in Visual Semantic AI (presented at ACM FAccT 2022).

December 2022 **MIT Tech Review**. The viral AI avatar app Lensa undressed me – without my consent. By Melissa Heikkila.
Refers to Markedness in Visual Semantic AI (presented at ACM FAccT 2022).

Academic Service

2025 AAAI 2025. **Program Committee**.

2024 AIES 2024. **Program Committee**.

AAAI 2024. **Program Committee**.

Association for Computational Linguistics Rolling Review 2024. Reviewer.

NeurIPS 2024. Reviewer.

2023 AAAI 2023. **Program Committee**.

NeurIPS 2023. Reviewer.

Nature Humanities and Social Science. Reviewer.

2022 ACM FAccT 2022. **Program Committee**.

ICML 2022. Reviewer.

NeurIPS 2022. Reviewer.

AIES 2022. Reviewer.

Industry experience

Winter 2023 **Deep Learning Research Intern, AKASA** – San Francisco, CA (Remote)
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.

2014-2021 **DisputeSoft: Software Dispute Experts** – Potomac, MD
Director of HR, Operations, and Marketing. Managed a small staff, and wrote memos to support technical analysis in complex software consulting matters.

Community Involvement

2023-2024 **University Children's Development School, Seattle, WA.**
Developed a system for allowing children to interact with multimodal AI systems and reason about AI fairness in a controlled context. System used as part of the curriculum for upper-elementary school children.

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

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

2016 **Presidential Election Judge, Montgomery County, MD.**
Assisted in facilitating voting in 2016 presidential election.

Lab Memberships

2023 – Present Volitional AI Lab (Howe Lab). UW iSchool.

2022 – Present User Empowerment Lab (Hiniker Lab). UW iSchool.

Professional Memberships

2022 – Present Association for Computing Machinery (ACM)

Association for the Advancement of Artificial Intelligence (AAAI)

Institute of Electrical and Electronics Engineers (IEEE)

2021 – Present Association for Computational Linguistics (ACL)

Technical Background

Programming	Python, R, C++, Java, C, Julia, PHP, Javascript, CSS/HTML
Libraries	PyTorch, Tensorflow, Transformers, Diffusers, Datasets, TRL, 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 • Critical and Cultural Perspectives on Information Science