Neil John Fasching

neilfasching@gmail.com · neilfasching.com
3620 Walnut St. Philadelphia, PA 19104

COMPUTATIONAL SOCIAL SCIENTIST

I am a computational social scientist currently completing my PhD at the University of Pennsylvania. My doctoral dissertation explores AI bias and evaluation in LLM-based research tools. Specifically, my work focuses on auditing these models for fairness and accuracy, developing standardized evaluation protocols, and validating outputs against human benchmarks. As a case study, I analyze over 100,000 podcast episodes, utilizing them to examine how these models process diverse speech patterns and cultural nuances—conditions where bias is most likely to emerge and affect computational social science findings.

EDUCATION

PhD -- Computational Social Science **Expected 2025**
University of Pennsylvania
Committee: Dr. Yphtach Lelkes, Dr. Duncan J. Watts, Dr. Sandra González-Bailón

Master's -- Statistics and Data Science **2023**
The Wharton School at the University of Pennsylvania

Master's -- Communication Science **2021**
University of Amsterdam

Bachelor's -- Political Science and Psychology **2014**
UC Santa Barbara

RECENT PROFESSIONAL EXPERIENCE

The Wharton School at the University of Pennsylvania **Jul 2022 - Present**
Co-Teacher for Statistics and Data Science Course
I teach Modern Data Mining, a PhD-level Data Science course in the Statistics department at Wharton. The course brings in a large set of cutting-edge machine learning techniques, such as Boosted Trees, CNNs, RNNs, and LLMs, combined with up-to-date case studies.

The University of Pennsylvania **Sep 2021 - Present**
Computational Research Fellow
Research Fellow collaborating with several professors, including Dr. Yphtach Lelkes and Dr. Duncan J. Watts, on projects that explore the influence of news media and social media on human behavior. My work involves employing diverse data collection, data mining, and analysis techniques, with a particular emphasis on leveraging large language models (LLMs). One example of this research is available at: Media Bias Detector.

AI AND DATA-DRIVEN PROJECTS

"Auditing the Algorithm: Evaluating Bias and Fairness in LLM Applications for Computational Social Science" (Dissertation)

My dissertation investigates AI bias and evaluation in computational social science, developing frameworks to enhance model transparency, fairness, and accountability. It focuses on auditing LLMs for representational harms, standardizing evaluation metrics to detect bias across contexts, and validating outputs against diverse benchmark datasets. By analyzing over 100,000 podcast episodes, I assess how models like Whisper perpetuate speech recognition disparities across demographics, how moderation endpoints exhibit cultural biases in content flagging, and how embedding models encode and amplify social stereotypes in classification tasks.

"Model-Dependent Moderation: Inconsistencies in Hate Speech Detection Across LLM-based Systems" (ACL 2025)

Led rigorous evaluation of AI bias in hate speech detection, systematically assessing seven leading content moderation platforms and revealing significant inconsistencies in classification outcomes for identical content across demographic groups. Developed novel experimental methodology using 1.3 million factorial-design sentences to quantify how model selection fundamentally determines content filtering decisions. Research demonstrated that moderation systems employ variable decision boundaries and detection thresholds across demographic categories, establishing new evaluation metrics for fairness assessment in AI content moderation systems.

"Toxic Air in the Public Square: Quantifying Toxicity of Political Discourse on Twitter" (EMNLP – Under Review 2025)

Conducted multi-model evaluation of AI toxicity detection in political discourse, analyzing 46.7 million tweets (2012-2022) across two state-of-the-art moderation platforms (OpenAI OMNI-MODERATION and Mistral MODERATION). Research demonstrated political tweets consistently showed higher toxicity than random content, with harassment metrics sharply increasing during 2016-2020. Developed novel comparative methodology to disaggregate political toxicity from broader platform trends, revealing distinct relationships between different toxicity types and account reach. Analysis quantified how content moderation systems detected varying toxicity patterns across demographic categories and political events, establishing new evaluation metrics for assessing AI moderation systems in democratic discourse.

"Automated Annotation with Generative AI Requires Validation" (arXic.org)

Evaluated AI annotation systems in computational social science through systematic assessment of LLM performance in content classification tasks. Developed human-in-the-loop feedback framework that iteratively improves machine annotation accuracy through targeted validation protocols. Quantified reliability metrics, bias patterns, and efficiency gains between GPT models and human coders, establishing benchmarks for responsible AI integration in research workflows. Demonstrated how validated LLM systems can produce accurate, reproducible classifications while reducing annotation costs and addressing fairness concerns in automated content analysis.

"Persistent Polarization: The Unexpected Durability of Political Animosity Around US Elections" (Science Advances)

Harnessing the power of different causal time-series analyses, I investigate the time-related characteristics of an election that influence support for political violence, political norm violations, and affective polarization. I then use these models to predict support for political violence in the buildup to a national political election.

PUBLICATIONS

Fasching, N. and Lelkes, Y. (2025). Model-dependent moderation: Inconsistencies in hate speech detection across LLM-based systems. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Fasching, N., Wolken, S., & Dörr, T. (2025). Toxic Air in the Public Square: Quantifying Toxicity of Political Discourse on Twitter. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. Under Review.

Fasching, N., Iyengar, S., Lelkes, Y., & Westwood, S. J. (2024). Persistent polarization: The unexpected durability of political animosity around US elections. *Science Advances*, 10(36), eadm9198.

Fasching, N., Arceneaux, K., & Bakker, B. N. (2024). Inconsistent and very weak evidence for a direct association between childhood personality and adult ideology. *Journal of Personality*, 92(4), 1100-1114.

Arceneaux, K., Bakker, B. N., **Fasching, N**., & Lelkes, Y. (2024). A critical evaluation and research agenda for the study of psychological dispositions and political attitudes. *Political Psychology*.

Schumacher, G., Homan, M. D., Rebasso, I., **Fasching, N.,** Bakker, B. N., & Rooduijn, M. (2024). Establishing the validity and robustness of facial electromyography measures for political science. *Politics and the Life Sciences*, 1-18.

Fasching, N., & Lelkes, Y. (2024). Ancestral Kinship and the Origins of Ideology. *British Journal of Political Science*, 54(1), 1-21.

Pangakis, N., Wolken, S., & **Fasching, N**. (2023). Automated annotation with generative AI requires validation. arXiv. *arXiv preprint arXiv:2306.00176*.

Bakker, B. N., Jaidka, K., Dörr, T., **Fasching, N.,** & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, 71(5), 715-738.

SKILLS

LLM Evaluation Techniques

- Red-teaming
- Bias Detection and Mitigation
- Workflow Standardization
- Human-in-the-Loop Validation
- Reproducibility
- Ethics and Compliance

Languages

- R (Expert)
- Python (Expert)
- SQL (Proficient)

Skills

- Posit Workbench
- Git/Github
- Amazon Web Services
- Microsoft Azure
- Google Colab
