Neil Fasching

Computational Social Scientist | AI Bias Researcher | PhD Candidate
neilfasching@gmail.com | neilfasching.com | 3620 Walnut St. Philadelphia, PA 19104

PROFESSIONAL SUMMARY

PhD candidate at the University of Pennsylvania with expertise in large-scale data analysis and AI Bias Research. Published in ACL and Science Advances. Led comprehensive evaluation of bias in hate speech detection across seven LLM-based content moderation systems. Created pipelines for analyzing tens of thousands of podcasts using multiple AI models, with emphasis on assessing model bias and fairness.

EDUCATION

PhD, Computational Social Science Expected 2025/2026
University of Pennsylvania

Master of Science, Statistics and Data Science 2023
The Wharton School, University of Pennsylvania

Master of Science, Communication Science 2021
University of Amsterdam

AI MODEL RESEARCH PROJECTS

Model-Dependent Moderation: Inconsistencies in Hate Speech Detection Across LLM-based Systems (Link to ACL Paper)

- Identified inconsistencies in how seven LLM-based systems (including Claude) handle harmful content
- Created a synthetic dataset of over 1.3 million sentences using a full factorial design to systematically test where models fail or exhibit bias, informing intervention strategies
- Quantified model inconsistencies in content filtering decisions by different demographic groups
- Established new evaluation metrics for model inconsistencies in AI systems

Leveraging Large Language Models to Evaluate Topics of Discussion, Misinformation, and Toxicity on Political Podcasts (Dissertation)

- Analyzed over 28,000 podcast episodes using several AI models for topics, misinformation, and toxicity
- Developed two novel frameworks for assessing the prevalence of misinformation and hate speech at scale
- Assessed LLM-based model performance and checked for bias and fairness in transcription, topic segmentation, misinformation identification and hate speech classification

Automated annotation with generative AI requires validation (Link to Paper)

- Developed and validated a 5-step workflow for LLM text annotation with human-in-the-loop validation
- Replicated 27 annotation tasks across 11 social science datasets, classifying over 200,000 text samples
- Introduced novel "consistency score" metric to identify edge cases and improve annotation reliability
- Demonstrated high performance variability across tasks (F1: 0.06-0.97), establishing need for validation

Toxic Air in the Public Square: Quantifying Toxicity of Political Discourse on Twitter (Link)

- Analyzed 46.7 million tweets (2012-2022) for toxicity (including harassing, hate, and violent speech)
- Developed scalable pipelines for toxicity measurement using OpenAI and Mistral moderation systems
- Utilized advanced ML models to analyze difference in toxicity across time, demographics, and topics

PROFESSIONAL EXPERIENCE

Computational Research Fellow **Sep 2021 - Present**
University of Pennsylvania

- Analyze large-scale datasets to study the trends, patterns, and effects of news media and social media
- Employ diverse data collection, data mining, and analysis techniques with emphasis on LLMs
- Develop novel pipelines for classifying unstructured text while minimizing bias
- One example: (mediabiasdetector.seas.upenn.edu/)

Co-Teacher, Modern Data Mining (PhD Level) **Jul 2022 - Present**
The Wharton School, University of Pennsylvania

- Teach PhD-level Data Science course in the Data Science department at Wharton
- Cover cutting-edge machine learning techniques including Boosted Trees, CNNs, RNNs, and LLMs
- Incorporate up-to-date case studies combining statistical theory with practical applications

TECHNICAL SKILLS

Programming Languages: Python, R, SQL, JavaScript

Machine Learning Libraries: PyTorch, TensorFlow, Keras, Hugging Face Transformers, scikit-learn, XGBoost, spaCy, NLTK, statsmodels

LLM/AI Tools: Large Language Models (OpenAI GPTs, Anthropic Claude, Google Gemini, Mistral Large); Content Moderation APIs (OpenAI, Mistral); Text Embeddings (OpenAI, Mistral, Google); Speech-to-Text (OpenAI Whisper), and many more

Data Processing: PySpark, PyArrow, Pandas, NumPy, Dask, dplyr/tidyverse, SparkR, arrow

Statistical Methods: Regression (Linear, Logistic, Multilevel), Neural Networks, Ensemble Methods, Time-Series Analysis, Causal Inference, Experimental Design

Platforms: AWS, Microsoft Azure, Google Colab, Posit Workbench, Git/GitHub

SELECT 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*

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

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

Fasching, N., Wolken, S., and Dörr, T. (2025). **Toxic Air in the Public Square: Quantifying Toxicity of Political Discourse on Twitter**. (Under Review).

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

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

RELEVANT EXPERIENCE FOR AI MODEL WELFARE & SAFETY

- Proven ability to identify model inconsistencies that could lead to welfare harms across 7 LLM systems
- Designed interventions to mitigate risks from AI bias, improving reliability for vulnerable user groups
- Created validation frameworks and consistency metrics that catch model failures before they cause harm
- Developed metrics for assessing when AI systems produce unreliable or potentially harmful outputs
- Experience translating abstract AI safety concerns into concrete, testable experiments
- Built human-in-the-loop systems to improve model trustworthiness and reduce bias