Terrence Neumann

Austin, TX 78746
(269) 370-8123 • tdneuman@gmail.com • terryneumann

About Me

I am a fifth-year PhD candidate researching responsible AI, developing methods to ensure large language models are safe, transparent, and socially beneficial. My work bridges machine learning, mechanistic interpretability, and computational social science to address pressing challenges at the intersection of AI and society.

Education

University of Texas at Austin **Austin, TX**
PhD – Information Systems, Ethical AI Emphasis *Expected 2026*
Thesis: Sociotechnical Controls for Mitigating AI Risk in the Absence of Ground Truth

Northwestern University **Evanston, IL**
MS in Analytics *December 2016*

Indiana University, Bloomington **Bloomington, IN**
BA in Economics and Mathematics (*Departmental Honors*) *May 2015*

Professional & Research Experience

SnorkelAI **Redwood City, CA**
Expert Contributor - Statistical and Logical Reasoning *January 2025 - present*
- Developed "reasoning traces" to complex (graduate-level causal reasoning and statistics) prompts so future LLMs learn necessary reasoning steps to solve complex problems.

University of Texas at Austin **Austin, TX**
Research Assistant – Good Systems Initiative *July 2021 - Present*
- Led research initiative investigating interpretable LLM steering towards certain social behaviors with sparse autoencoders (SAEs). *Publication forthcoming*.
- Developed statistical methodologies for automatic and low-cost quality checks for LLM-simulated opinion surveys. *Publication forthcoming*.
- Led multiple research projects related to ethical and responsible use of AI in fact-checking workflows, especially related to prioritizing what claims to fact-check. Published at ACM FAccT '22, '23.

Crime and Education Labs, University of Chicago **Chicago, IL**
Data Scientist, Senior Data Scientist *Jan 2017 – June 2021*
- Designed and deployed person-based predictive models using City of Chicago administrative data to identify at-risk individuals to be targeted with additional social services, especially related to domestic violence and education dropout.
- Developed novel causal inference technique (`ElasticSynth`) to determine effect of change in policing management on crime levels in Chicago.

Technical Skills

Programming Languages & ML Frameworks: Python, R, PyTorch, scikit-learn, LangChain, MCP
Large Language Models: Fine-tuning, In-Context Learning, Knowledge Injection, Prompt Engineering
LLM Safety & Evaluation: Sparse auto-encoders, Statistical approaches to LLM evaluation
Causal Inference: Hypothesis Testing, Matching, Synthetic Controls, Bootstrapping, Permutation Tests

Relevant Publications

- **Terrence Neumann** and Nicholas Wolczynski. *Does AI-Assisted Fact-Checking Disproportionately Benefit Majority Groups Online?* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT). 2023.
- Tanriverdi, Hüseyin, John-Patrick Akinyemi, and **Terrence Neumann**. *Mitigating Bias in Organizational Development and Use of Artificial Intelligence.* Proceedings of the 2023 International Conference on Information Systems (ICIS). (2023).
- **Terrence Neumann**, Maria De-Arteaga, and Sina Fazelpour. *Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT). 2022.

Working Papers

- **Terrence Neumann** and Yan Leng. *From Statistical Patterns Emerge Human-Like Behaviors: How LLMs Learn Social Preferences*. Under Review. 2025.
- **Terrence Neumann**, Maria De-Arteaga, and Sina Fazelpour. *Should You Use LLMs to Simulate Opinions? Quality Checks for Early-Stage Deliberation.* Under Review. 2025.
- **Terrence Neumann**, Sooyong Lee, Maria De-Arteaga, Sina Fazelpour, and Matthew Lease. *Diverse, but Divisive. LLMs Can Exaggerate Gender Differences in Opinion Related to Harms of Misinformation. arxiv*. 2024.
- **Terrence Neumann**, Maria De-Arteaga, Sina Fazelpour, Maytal Saar-Tsechansky, and Matthew Lease. *Informational Justice in AI-Assisted Fact-Checking.* SSRN. 2024.
- **Terrence Neumann** and Bryan Jones. *PRISM: A Design Framework for Open-Source Foundation Model Safety. arxiv*. 2024.
- Max Kapustin, **Terrence Neumann**, and Jens Ludwig. Policing and management. No. w29851. National Bureau of Economic Research, 2022.

Fellowship/Awards

Brumley Fellowship in Cybsersecurity. 2023-2024: Investigating the risks associated with open-sourcing advanced AI technologies, like LLMs.

NYU Policing Project Fellow. 2023-2025: Providing data assistance and thought leadership on "Reimagining Public Safety" project, which aims to reduce violent interactions between police and civilians via alternative response.

Conference Presentations

Wharton AI & the Future of Work Conference: 2025

Conference on Information Systems and Technology (CIST): 2024

INFORMS Annual Meeting: 2024

ACM Conference of Fairness Accountability and Transparency (FAccT): 2022, 2023

Service

ACM WWW (The Web Conference): 2025 - Reviewer

AAAI/ACM Conference on AI, Ethics, and Society: 2024, 2025 - Program Committee

Conference on Information Systems and Technologies (CIST): 2024, 2025 - Program Committee

INFORMS Data Science Workshop: 2024, 2025 - Program Committee