Peter Kirgis

42 Loyola Terrace, San Francisco 94117
pk7019@princeton.edu — peterkirgis.github.io

Education

Princeton School of Public and International Affairs *May 2025*
Master of Public Affairs (MPA), Certificate in Statistics and Machine Learning
- Cumulative GPA: 3.94

Williams College *June 2020*
Bachelor of Arts with Honors in Political Economy and Philosophy

Research

Towards a Science of AI Agent Reliability *2026*
Rabanser, S., Kapoor, S., **Kirgis, P.**, Liu, K., Utpala, S., & Narayanan, A.
arXiv preprint arXiv:2602.16666
- Proposed twelve metrics decomposing AI agent reliability along four dimensions (consistency, robustness, predictability, safety) and evaluated 14 models, finding that recent capability gains have yielded only small improvements in reliability. Paper | Website

LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces *2026*
Kirgis*, P., Hawriluk*, B., Feng, S., Bilimer, A., Paech, S., & Tufekci, Z. (* Equal contribution)
Accepted at IASEAI 2026
- Conducted an audit study comparing LLM behavior across API and chat interfaces, documenting large differences in sycophancy, escalation, and delusion reinforcement between environments and across model versions. Website

Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation *2025*
Kapoor, S., Stroebl, B., **Kirgis, P.**, Nadgir, N., Siegel, Z. S., Wei, B., ... & Narayanan, A.
Accepted at ICLR 2026
- Executed large-scale data analysis of 21,000+ AI agent rollouts to evaluate the capabilities and limitations of frontier AI agents. Paper | Code

Differences in the Moral Foundations of Large Language Models *2025*
Kirgis, Peter
arXiv preprint arXiv:2511.11790
- Administered synthetic experiments to sixteen frontier large language models to elicit diverse moral judgments. Used PCA and NLP methods to explore and visualize bias and clustering of LLM responses relative to a human baseline. Paper | Code | Slides

Desegregation Paradox? A Model of LIHTC's Effects on Economic Segregation *2024*
Kirgis, Peter and Teddy Knox
Preprint
- Used novel data sources to simulate the effect of redistributing LIHTC residents across US metropolitan areas and observed that LIHTC's presence is associated with a decrease in economic segregation. Paper

Invited Talks

"A Few Insights from the Analysis of Over 2,000 AI Agent Logs" NeurIPS AI Evaluator Forum, *Dec 2025*

"Is Consciousness Prerequisite for Moral Patienthood?" ELEOS AI Consciousness Conference, *Nov 2025*

"Differences in the Moral Foundations of Large Language Models" PICSciE/CSML Colloquium, *Apr 2025*

"Improving Imputation Accuracy" Coding it Forward Fellowship Demo Day, *Aug 2024*

Professional Experience

Center for Information Technology Policy, Princeton University *August 2025 – Present*
Research Scientist

- Working with Arvind Narayanan and Zeynep Tufekci on AI agent evaluation and the societal impacts of AI.
- Contributing to the development of an AI agent reliability index with the Holistic Agent Leaderboard (HAL) team.
- Led an audit study of sycophancy and delusion reinforcement in ChatGPT, accepted at IASEAI 2026.
- Leading a multi-organization project defining threats to credible AI agent evaluation.
- Designed and implemented a data pipeline to extract and grade 100GB of agent logs using Python.

SPAR (Supervised Program for Alignment Research) *January 2026 – Present*
Research Intern (Part-Time)

- Working with Tim Hua to study how frontier LLMs iteratively refine their own constitutions, model specs, and system prompts.

United States Census Bureau (Coding it Forward) *June 2024 – August 2024*
Data Scientist Fellow

- Built sorting algorithms and k-nearest neighbors regression models in Python to improve existing imputation methods within the Economic Statistical Methods Division.
- Designed Python data visualizations comparing imputation strategies across several loss functions.

Commonwealth of Massachusetts *November 2021 – July 2023*
Data Analyst

- Led inter-agency data sharing and program evaluation efforts on behalf of the Chief Data Officer.
- Designed and maintained PostgreSQL ETL pipelines and automated data workflows with Python, AWS Lambda, and Tableau.
- Implemented the Commonwealth's first differential privacy algorithm on longitudinal datasets related to early childhood education and workforce development.