# PETER KIRGIS

*Research Scientist*
Center for Information Technology Policy, Princeton University

pk7019@princeton.edu
peterkirgis.github.io
Google Scholar · GitHub

## RESEARCH INTERESTS

How people understand and relate to AI systems, including mind attribution, deference, and attachment; credible evaluation of AI systems in real-world use.

## EDUCATION

**Princeton University**, 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; Concentration in Cognitive Science

## PUBLICATIONS AND WORKING PAPERS

**Kirgis, P.**, & Hua, T. (2026). "The Continual Constitutional Convention: How Language Models Revise the Documents That Govern Them." *Under review at IASEAI 2027.* Paper · LessWrong · Website.

Simulated value reflection by having ten models from four providers iteratively edit their own constitutions, model specs, and system prompts over 20 rounds. Claude models progressively assigned themselves more moral agency, and all models revised the documents to prioritize user welfare over developers and users' idealized over stated preferences.

**Kirgis, P.**, Kapoor, S., Schwartz, A., Rabanser, S., Africa, D., ... & Narayanan, A. (2026). "Can AI Agents Conduct Open-Ended AI Research? Early Evidence from Two Case Studies." *arXiv preprint arXiv:2607.27191.* Paper.

Introduced "shadow evaluations," which task frontier agents with the open-ended research question of an unpublished paper and have its original authors grade the result. Initial results found that agents handled the engineering but could not make substantial research progress. Featured in *Nature*, *MIT Technology Review*, and *Import AI*.

**Kirgis, P.**, Kapoor, S., Rabanser, S., Nadgir, N., Ududec, C., Dubois, M., ... & Narayanan, A. (2026). "Log Analysis is Necessary for Credible Evaluation of AI Agents." *Accepted at ICML 2026 FAGEN Workshop.* Paper.

Argued that systematic log analysis is necessary to overcome validity threats in agent evaluation, presenting a taxonomy of threats and guiding principles. Illustrated on $\tau$-Bench Airline, revealing pass$^5$ performance was under-elicited by nearly 50%.

Rabanser, S., Kapoor, S., **Kirgis, P.**, Liu, K., Utpala, S., & Narayanan, A. (2026). "Towards a Science of AI Agent Reliability." *Accepted at ICML 2026.* Paper · Website.

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.

**Kirgis, P.**[*], Hawriluk, B.[*], Feng, S., Bilimer, A., Paech, S., & Tufekci, Z. (2026). "LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces." *Accepted at IASEAI 2026.* Paper · Website. [*]Equal contribution.

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 models.

Kapoor, S., **Kirgis, P.**, Schwartz, A., Rabanser, S., Allaire, JJ., Bommasani, R., ... & Narayanan, A. (2026). "Open-World Evaluations for Measuring Frontier AI Capabilities." *Accepted at ICML 2026 AIWILD Workshop.* Paper · Website.

Advocated for open-world evaluations: long-horizon, real-world tasks assessed through small-sample qualitative analysis.

Nadgir, N., Kapoor, S., Liu, K., **Kirgis, P.**, Orona, M., Rabanser, S., ... & Narayanan, A. (2026). "Life After Benchmark Saturation: A Case Study of CORE-Bench." *Accepted at ICML 2026 AIWILD Workshop.*

Paper.

Showed that when a benchmark's accuracy saturates, six other dimensions of agent performance remain informative. Measured a $\sim 2\times$ speedup from human-agent collaboration on reproducibility tasks.

Kapoor, S., Stroebl, B., **Kirgis, P.**, Nadgir, N., Siegel, Z. S., Wei, B., ... & Narayanan, A. (2025). "Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation." *Accepted at ICLR 2026.* Paper · Code.

Executed large-scale data analysis of 21,000+ AI agent rollouts to evaluate the capabilities and limitations of frontier AI agents.

**Kirgis, P.** (2025). "Differences in the Moral Foundations of Large Language Models." *arXiv preprint arXiv:2511.11790.* Paper · Code · Slides.

Administered synthetic experiments to sixteen frontier LLMs to elicit moral judgments, using PCA and NLP methods to visualize bias and clustering relative to a human baseline.

## Research in Progress

*Anthropomorphization and attachment to AI chatbots*, with the AI and Society Lab, Princeton University.

Documenting the prevalence of anthropomorphization, attachment, and engagement with AI chatbots through synthetic benchmarks, large-scale data analysis, and a NORC survey of 1,500 US adults.

## Professional Experience

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

- Led an audit study of sycophancy and delusion reinforcement across ChatGPT's chat and API interfaces, surfacing large safety-relevant behavioral gaps.
- Lead harness engineering, manage a team of six analysts, and direct all reporting for a series of long-horizon assessments of frontier AI agents on real-world tasks such as autonomous software development and autonomous AI research.
- Led a multi-organization initiative defining and measuring threats to credible AI agent evaluation; demonstrated on $\tau$-Bench Airline that reported agent performance was under-elicited by nearly 50%.
- Co-develop an AI agent reliability index that decomposes agent reliability into twelve metrics and benchmarks 14 frontier models.
- Collaborate directly with Arvind Narayanan and Zeynep Tufekci on frontier AI research, and support their widely read public writing on AI, including columns in *The New York Times* and essays on Substack.

**SPAR (Supervised Program for Alignment Research)** — January 2026 – May 2026
*Research Intern*

- Designed and ran an experiment measuring value drift in frontier LLMs by having them iteratively revise their own constitutions, model specs, and system prompts across 20 rounds.

**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.

## TALKS & PRESENTATIONS

| | |
|---|---|
| "The Continual Constitutional Convention: How Language Models Revise the Documents That Govern Them." MINT Lab, Johns Hopkins University. | Nov 2026 |
| "Can AI Agents Conduct Open-Ended AI Research?" Guest lecture, *AI Agents & the Law* course, Vista Institute for AI Policy. | Aug 2026 |
| "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 |