

# Hsiang Hsu

RESEARCH SCIENTIST — Frontier Model Evals, Alignment Robustness, and Trustworthy AI

✉ [hsiang.s.hsu@gmail.com](mailto:hsiang.s.hsu@gmail.com) |  Google Scholar |  LinkedIn

## Summary

---

Research Scientist working on frontier model evaluation, alignment robustness, and trustworthy AI. My research develops stress tests for hidden failure modes that standard average-case evaluations miss, including residual knowledge after unlearning, prompt-specific reward tail risk, and behavioral arbitrariness among near-equivalent models. My work turns uncertainty, privacy, and robustness failures into operational evals for model behavior, alignment, and intervention verification. My research has appeared at NeurIPS, ICLR, ICML, TPAMI, and TIT, including a NeurIPS Oral, NeurIPS Spotlight, ICML Spotlight, and the Meta PhD Fellowship.

## Research Artifacts

---

- AlignmentTailBench** Ranks high-risk prompts missed by average reward metrics.
- ResidualKnowledgeProbe** Detects residual knowledge recovered under perturbations after unlearning.
- BehavioralMultiplicity** Surfaces arbitrary decisions among similarly performing models.

## Professional Experience

---

2023-2026 **Research Lead, Applied Intelligence Lab, JPMorgan Chase & Co.**, New York, NY, USA

- Led research on **adversarial evaluation of machine unlearning**, showing that models can appear to forget under standard tests while retaining **residual knowledge** under perturbed deletion queries; resulted in a NeurIPS 2025 publication and follow-up protocols for **intervention verification**, stress-testing whether forgetting claims remain valid under distribution shift.
- Developed methods for **failure-mode discovery and provenance analysis in generative systems**, including **low-entropy LLM watermarking**, partial-LLM text detection, and perturbation-based hallucination probing, exposing brittle behaviors missed by aggregate evaluation metrics and turning them into measurable signals for model debugging.
- Built an evaluation agenda around **reward and decision instability**, studying prompt-specific reward distributions, tail risks in inference-time alignment, and **predictive multiplicity** among near-equivalent models; framed these phenomena as operational evals for identifying prompts, checkpoints, and model behaviors that are unstable despite strong average performance.
- Advanced **model safety and governance** frameworks without sensitive attributes, using **predictive multiplicity** and **decision uncertainty** as evaluation signals to surface unstable decisions, identify hidden behavioral subgroups, and improve fairness when demographic labels are unavailable or restricted.

2017-2023 **Research Assistant, Harvard University**, Cambridge, MA, USA

- Developed **information-theoretic methods for model behavior beyond aggregate accuracy**, including fairness interventions, privacy leakage control, and representation-level attribution, establishing a foundation for evaluating how models trade off accuracy, robustness, privacy, and individual-level reliability.
- Introduced frameworks for **predictive multiplicity**, **behavioral arbitrariness**, and the **Rashomon effect**, showing that near-equivalent models can produce conflicting individual decisions despite similar aggregate performance; resulted in NeurIPS 2022 and NeurIPS Spotlight 2023 publications and research recognized by the **Meta PhD Fellowship**.
- Designed fairness interventions based on information projection for multi-class prediction and analyzed how group-level constraints can induce individual-level arbitrariness, resulting in NeurIPS Oral and Spotlight presentations.

2021 **ML Research Intern, Apple**, Health AI Team, Cupertino, CA, USA

- Developed hybrid expert-model augmentation methods for physiological forecasting, improving out-of-distribution generalization by combining mechanistic ODE-based simulators with learned models.

2019-2020 **Research Intern, Pinterest & Salesforce IQ**, Palo Alto, CA, USA

## Education

---

### Harvard University

PHD & MS IN COMPUTER SCIENCE

Cambridge, MA, USA

Sept. 2017 - May 2023

- Thesis: Information-Theoretic Tools for Machine Learning Beyond Accuracy — fairness, privacy, and decision uncertainty.

### National Taiwan University

MS IN ELECTRICAL ENGINEERING & BS IN ELECTRICAL ENGINEERING AND MATHEMATICS

Taipei, Taiwan

Sept. 2014 - June 2016

- Thesis: Efficient Resource Allocation on Graphs - Crowdsourcing.
- Best Master's Thesis Award (1st/107, school-wide) and National Young Scholar Best Paper Award (2nd).

## Professional Recognition & Service

---

### RESEARCH RECOGNITION

- 2021 **Meta PhD Fellowship in Applied Statistics**: Selected as one of 21 senior PhD candidates worldwide from 2,000+ applicants.
- 2022-2025 **Top Conference Paper Distinctions**: NeurIPS Oral, NeurIPS Spotlight, and ICML Spotlight for work on fair prediction, individual arbitrariness, and private attribute protection.
- 2024 **JPMC Prolific Inventor**: Awarded for filing 10+ U.S. patents in Trustworthy AI, model safety, and governance.

## SERVICE

- 2021–2025 **Area Chair & Program Chair:** ITR3 Workshop @ ICML 2021, ACM FAccT 2025.  
2025–2026 **Tutorial Organization:** ACM FAccT 2025, AAAI 2026.  
2017–2026 **Reviewer:** ICML, NeurIPS, AISTATS, ICLR, ISIT, ITW, FAccT, TIT, TMLR, TheWebConf.

## SELECTED MEDIA

- 2025 **The Daily Upside:** JPMC machine unlearning and privacy guardrails.  
2024 **UT News, The Daily Texan, PetaPixel:** Generative machine unlearning.  
2022 **Meta Research Spotlight:** Rashomon effect and predictive multiplicity.

## Selected Publications

---

### FRONTIER MODEL EVALS, UNLEARNING, AND GENERATIVE MODEL SAFETY

1. **Hsiang Hsu**, Pradeep Niroula, Zichang He, Ivan Brugere, Freddy Lecue, Chun-Fu Chen. “The Unseen Threat: Residual Knowledge in Machine Unlearning under Perturbed Samples”. In Advances in Neural Information Processing Systems: **NeurIPS 2025**.
2. Dor Tsur, Carol Xuan Long, ..., **Hsiang Hsu**, Chun-Fu Chen, Haim H Permuter, Flavio Calmon. “HeavyWater and SimplexWater: Distortion-Free LLM Watermarks for Low-Entropy Next-Token Predictions”. In Advances in Neural Information Processing Systems: **NeurIPS 2025**.
3. Eric Lei, **Hsiang Hsu**, Chun-Fu Chen. “PaLD: Detection of Text Partially Written by Large Language Models”. In Proceedings of the International Conference on Learning Representations: **ICLR 2025**.
4. Seongmin Lee, **Hsiang Hsu**, Chun-Fu Chen, Duen Horng Chau. “Probing LLM Hallucination from Within: Perturbation-Driven Approach via Internal Knowledge”. In Proceedings of the IEEE International Conference on Big Data: **IEEE BigData 2025**.
5. Guihong Li, **Hsiang Hsu**, Chun-Fu Chen, Radu Marculescu. “Machine Unlearning for Image-to-Image Generative Models”. In Proceedings of the International Conference on Learning Representations: **ICLR 2024**.
6. Guihong Li, **Hsiang Hsu**, Chun-Fu Chen, Radu Marculescu. “Fast-NTK: Parameter-Efficient Unlearning for Large-Scale Models”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: **CVPR 2024**.

### PREDICTIVE MULTIPLICITY, DECISION UNCERTAINTY, AND ALIGNMENT ROBUSTNESS

7. **Hsiang Hsu**, Ivan Brugere, Shubham Sharma, Freddy Lecue, Chun-Fu Chen. “RashomonGB: Analyzing the Rashomon Effect and Mitigating Predictive Multiplicity in Gradient Boosting”. In Advances in Neural Information Processing Systems: **NeurIPS 2024**.
8. **Hsiang Hsu**, Guihong Li, Shaohan Hu, Chun-Fu Chen. “Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation”. In Proceedings of the International Conference on Learning Representations: **ICLR 2024**.
9. Carol Long, **Hsiang Hsu**, Wael Alghamdi, Flavio Calmon. “Individual Arbitrariness and Group Fairness”. In Advances in Neural Information Processing Systems: **NeurIPS Spotlight 2023**.
10. Bogdan Kulynych, **Hsiang Hsu**, Carmela Troncoso, Flavio P Calmon. “Arbitrary Decisions are a Hidden Cost of Differentially Private Training”. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency: **FAccT 2023**.
11. **Hsiang Hsu**, Flavio P Calmon. “Rashomon Capacity: A Metric for Predictive Multiplicity in Classification”. In Advances in Neural Information Processing Systems: **NeurIPS 2022**.
12. Prakhar Ganesh, **Hsiang Hsu**, Golnoosh Farnadi. “Data as a Lever: A Neighbouring Datasets Perspective on Predictive Multiplicity”. arXiv preprint.
13. **Hsiang Hsu**, Eric Lei, Chun-Fu Chen. “Best-of-Tails: Bridging Optimism and Pessimism in Inference-Time Alignment”. arXiv preprint.

### FAIRNESS, PRIVACY, AND INFORMATION-THEORETIC ML

14. Yizhuo Chen, Chun-Fu Chen, **Hsiang Hsu**, Shaohan Hu, Tarek Abdelzaher. “PASS: Private Attributes Protection with Stochastic Data Substitution”. In Proceedings of the International Conference on Machine Learning: **ICML Spotlight 2025**.
15. Yizhuo Chen, Chun-Fu Chen, **Hsiang Hsu**, Shaohan Hu, Marco Pistoia, Tarek Abdelzaher. “MaSS: Multi-attribute Selective Suppression for Utility-preserving Data Transformation from an Information-theoretic Perspective”. In Proceedings of the International Conference on Machine Learning: **ICML 2024**.
16. Wael Alghamdi<sup>†</sup>, **Hsiang Hsu**<sup>†</sup>, Haewon Jeong, ..., Flavio Calmon. “Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection”. In Advances in Neural Information Processing Systems: **NeurIPS Oral 2022**.
17. **Hsiang Hsu**, Salman Salamatian, Flavio Calmon. “Generalizing Correspondence Analysis for Applications in Machine Learning”. In IEEE Transactions on Pattern Analysis and Machine Intelligence: **TPAMI 2021**.
18. Sungmin Cha, **Hsiang Hsu**, Taebaek Hwang, Flavio P Calmon, Taesup Moon. “CPR: Classifier-Projection Regularization for Continual Learning”. In Proceedings of the International Conference on Learning Representations: **ICLR 2021**.
19. **Hsiang Hsu**, Shahab Asoodeh, Flavio Calmon. “Information-Theoretic Privacy Watchdogs”. In Proceedings of the IEEE International Symposium on Information Theory: **ISIT 2019**.