Hsiang Hsu

Research Scientist — Knowledge Control, Alignment, and Model Reliability
hsiang.s.hsu@gmail.com | Hsiang Hsu | Scholar | GitHub | LinkedIn

Summary

Research Scientist focused on revealing failure modes in knowledge control for foundation models, spanning inference-time alignment—when models fail to reliably generate or select desirable responses from their existing capabilities—and knowledge removal—when targeted knowledge is entangled with retained knowledge and apparent forgetting reflects suppression rather than true removal. I develop evaluations that expose failures missed by standard benchmarks and translate them into practical monitoring and intervention methods. My work also studies knowledge uncertainty—how models with similar aggregate performance can rely on different internal representations or knowledge pathways, and how these micro-level differences can lead to divergent individual decisions and behaviors. 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 Portfolio

Inference-Time Alignment & Knowledge Steering
Developing methods to control how foundation models generate, explore, and select responses, including reward-tail-aware selection, in-context candidate generation for broader response coverage, and activation-level steering.

Knowledge Removal & Entanglement
Studying whether machine unlearning truly removes knowledge or merely suppresses it when forget and retain concepts share representations, using adversarial residual-knowledge audits, SAEs, and activation-level analysis.

Knowledge Uncertainty & Internal Pathways
Studying how near-equivalent models can achieve similar aggregate performance through different features, representations, or knowledge pathways, using predictive multiplicity and Rashomon analysis to expose unstable decisions and representation-level disagreement.

Professional Experience

2023-2026 **Research Lead, Applied Intelligence Lab, JPMorgan Chase & Co.,** New York, NY, USA
- Led research on **machine unlearning and residual knowledge**, using adversarial audits to reveal knowledge missed by standard forgetting tests and distinguish removal from suppression.
- Developed **inference-time alignment** methods addressing **prompt-specific reward variation and limited response coverage**, using reward-tail-aware adaptive selection and in-context candidate generation.
- Developed methods for **LLM reliability and provenance**, including low-entropy watermarking, partial-LLM text detection, and perturbation-based hallucination probing.

2017-2023 **Research Assistant, Harvard University,** Cambridge, MA, USA
- Developed **predictive multiplicity and decision-uncertainty** frameworks to reveal instability hidden by aggregate performance, showing how near-equivalent models can produce conflicting individual decisions and using disagreement as a signal for reliability and fairness.
- Showed that common interventions, including **fairness constraints and differential privacy**, can induce **individual-level uncertainty and behavioral arbitrariness** even when aggregate objectives improve, characterizing trade-offs between group-level guarantees and individual reliability.
- Developed **information-theoretic methods** for analyzing and controlling model behavior, including privacy leakage, representation-level attribution, and information-projection interventions for fair prediction.

2021 **ML Research Intern, Apple,** Health AI Team, Cupertino, CA, USA
- Addressed distribution-shift failures in physiological forecasting by developing hybrid expert-model augmentation that combines mechanistic ODE-based simulators with learned models to improve **OOD generalization**.

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

Education

Harvard University *Cambridge, MA, USA*
PhD & MS in Computer Science *Sept. 2017 - May 2023*
- Thesis: Information-Theoretic Tools for Machine Learning Beyond Accuracy — fairness, privacy, and decision uncertainty.

National Taiwan University *Taipei, Taiwan*
MS in Electrical Engineering & BS in Electrical Engineering and Mathematics *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, ACL, ISIT, ITW, FAccT, TIT, TMLR, TheWebConf; ICML Gold Reviewer 2026.

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

Knowledge Removal: Machine Unlearning and Residual Knowledge

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. 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**.
3. 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**.

Knowledge Uncertainty: Rashomon Effect, Multiplicity, and Decision Instability

4. **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**.
5. **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**.
6. Carol Long, **Hsiang Hsu**, Wael Alghamdi, Flavio Calmon. "Individual Arbitrariness and Group Fairness". In Advances in Neural Information Processing Systems: **NeurIPS Spotlight 2023**.
7. 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**.
8. **Hsiang Hsu**, Flavio P Calmon. "Rashomon Capacity: A Metric for Predictive Multiplicity in Classification". In Advances in Neural Information Processing Systems: **NeurIPS 2022**.

Inference-Time Alignment and Generative Model Reliability

9. **Hsiang Hsu**, Eric Lei, Chun-Fu Chen. "Best-of-Tails: Bridging Optimism and Pessimism in Inference-Time Alignment". arXiv preprint.
10. 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**.
11. 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**.
12. 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**.

Foundations: Fairness, Privacy, and Information-Theoretic ML

13. 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**.
14. 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**.
15. Wael Alghamdi†, **Hsiang Hsu**†, 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**.
16. **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**.
17. 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**.
18. **Hsiang Hsu**, Shahab Asoodeh, Flavio Calmon. "Information-Theoretic Privacy Watchdogs". In Proceedings of the IEEE International Symposium on Information Theory: **ISIT 2019**.