THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PhD Dissertation Presentation

Yuwei Cheng
Department of Statistics
The University of Chicago

"Aligning Machine Learning Systems with Human Preferences: Robustness, Personalization, and Evaluation"

May 7, 2026, at 2:30 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

This dissertation studies how to align machine learning systems with human preferences in complex, real-world environments. I approach this through three dimensions: robustness**,** personalization, and evaluation.

First, in robustness, I study learning under imperfect human feedback and establish an intrinsic efficiency–robustness trade-off: faster-converging algorithms are more vulnerable to corrupted or adversarial observations. This issue is amplified in multi-agent settings, where even a single compromised agent can influence system-wide behavior.

Second, for personalization, I develop contextual reinforcement learning frameworks for applications such as auto-bidding in digital advertising. The key challenge is learning from delayed and cumulative effects while adapting to heterogeneous users. I propose algorithms that balance exploration and exploitation and achieve near-optimal guarantees.

Finally, in evaluation, I introduce a principled, tree-based framework to characterize diversity in model outputs, particularly for large language models. I show that fine-tuning improves efficiency by redistributing uncertainty—focusing entropy on informative parts of the generation while reducing low-value continuations.
