STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Nan Cui
Degree:	Doctor of Philosophy
School/Department:	Charles V. Schaefer, Jr. School of Engineering & Science / Computer Science
Date:	Monday, August 18th, 2025
Time/Location:	2:00 p.m. / Gateway North 303
Title:	Fairness-Informed Machine Learning
Chairperson:	Dr. Yue Ning, Department of Computer Science, School of Engineering & Science
Committee Members:	Dr. Yue Ning, Department of Computer Science, School of Engineering & Science Dr. Wendy Hui Wang, Department of Computer Science, School of Engineering & Science Dr. Philippos Mordohai, Department of Computer Science, School of Engineering & Science Dr. Violet Chen, School of Business

ABSTRACT

As machine learning models are increasingly deployed in socially impactful domains, addressing algorithmic bias in model design and deployment has become a critical concern. This thesis presents a series of contributions aimed at mitigating bias to improve both individual- and group-level fairness across several learning paradigms, including active learning, federated learning, and large language models.

We begin by introducing a metric-based fairness control framework for active learning, with a new theoretical analysis of how to incorporate instance-level constraints during label-efficient training. This work highlights how biased sample selection can undermine model reliability and proposes a method to improve individual-level treatment consistency. Next, we examine distributed training settings and propose fairness-aware approaches for federated learning. We introduce a training strategy that combines focal loss and aggregation techniques to reduce disparities across data partitions in a federated graph neural network. We design structure-sensitive aggregation strategies to handle group disparities in non-IID graph data. In the last work, we investigated fairness issues in large language models (LLMs)-based recommender systems (RecLLMs). While LLMs enable dynamic, context-aware recommendations, they may inadvertently amplify user-side demographic unfairness. We propose a lightweight debiasing method that removes sensitive attribute information from LLM representations using a closed-form kernelized projection. To restore task utility, we design a gated Mixture-of-Experts adapter that selectively reintroduces non-sensitive, task-relevant information. Our approach effectively mitigates counterfactual leakage without requiring full model fine-tuning, offering an efficient and scalable fairness solution for RecLLMs. Together, these contributions offer practical tools and theoretical insights for reducing algorithmic bias in real-world machine learning systems. This work aims to support the development of more consistent, reliable, and socially responsible AI technologies.