STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate: Xuting Tang
Degree: Doctor of Philosophy
School/Department: Computer Science
Date: Monday, July 3rd, 2023
Time/Location: 10:00 AM – 11:30 AM EST. https://stevens.zoom.us/my/xujia
Title: Trust in the AI-R

Chairperson:

Dr. Jia Xu, Department of Computer Science, Stevens Institute of Technology

Committee Members:

Dr. Shusen Wang, Xiaohongshu, Xingin Information Technology Co Ltd

Dr. Steve Yang, School of Business, Stevens Institute of Technology

Dr. Enrique Dunn, Department of Computer Science, Stevens Institute of Technology

Dr. Samantha Kleinberg, Department of Computer Science, Stevens Institute of Technology

Dr. Richard Sharp, Cambridge Mobile Telematics

ABSTRACT

The call for trustworthy AI has been a long pursuit, yet still far to reach. My thesis joins the initiative that brings three trustworthiness aspects, Accuracy, Interpretability, and Resilience (AI-R), under one umbrella. Unlike conventional approaches striving to balance AI-R aspects by sacrificing model performance for interpretability or compromising transparency for resilience, we take a unique direction and show that model interpretation helps to improve prediction quality and model resilience. Our methods advance attention mechanisms, LIME, SHAP, self-learn post-hoc explainability models, and incorporate model-intrinsic techniques into a multi-agent reinforcement learning (MARL) framework. Extensive experiments demonstrate that our methods boost model AI-R over various tasks in computational linguistics, the Markov game, criminal analysis, and financial decisions, including (1) significant improvement in machine translation accuracy in large-scale; (2) speedup of the model adaptation time when new agents join in MARL competitive and collaborative games; (3) mitigating algorithmic bias in criminal charges and financial decision-making. This dissertation is among the works taking the first step in unifying previously unaligned AI-R perspectives into one goal for trustworthy AI.