Foundations of Trustworthy Sequential Decision-Making:
Privacy, Robustness and Fairness

Overview:

In an era where technology increasingly intersects with every aspect of our lives, the potential of sequential decision-making in machine learning (ML) unfolds as a transformative force. Systems like ChatGPT, through reinforcement learning (RL) and bandit algorithms, exemplify how technology can learn from and adapt to our individual behaviors, preferences, and needs, promising unparalleled personalization in healthcare, education, commerce, and beyond. However, this future is not without its challenges. As these intelligent systems become more integrated into our daily lives, the imperative for ensuring their trustworthiness—through safeguarding user privacy, ensuring system robustness, and promoting fairness—becomes increasingly critical. This CAREER project proposes to delve into these challenges, asking: *How can we build data-driven sequential decision-making systems that users can truly trust?* This project will tackle this question through a theoretical lens (via tools ranging from learning theory, optimization, differential privacy, robust statistics, and game theory), striving to establish fundamental theoretical limits and uncover innovative algorithmic principles for trustworthy interactive decision-making, from the perspectives of privacy, robustness, and fairness.

Keywords: Sequential decision-making; Bandit learning; Reinforcement learning; Differential privacy; Robustness; Fairness

Intellectual Merit:

This CAREER project marks a significant advancement in laying the groundwork for trustworthy sequential decision-making. It diverges from traditional focuses on statistical learning (e.g., supervised learning) to address the unique, intricate challenges of interactivity in machine learning (e.g., **bandits and RL**), such as the nuanced trade-offs between exploring new strategies and exploiting known information. This venture is structured around three core thrusts, each dedicated to a fundamental aspect of trustworthiness:

Privacy: The privacy thrust aims to revolutionize how we understand and implement differential privacy (DP) in contexts where learning agents and users interact dynamically. This includes establishing theoretical limits of DP in interactive systems, optimizing the trade-off between privacy, utility, and communication in federated learning environments, and devising a generalized framework for privacy-preserving interactive decision-making. These goals are ambitious, targeting fundamental and open questions about DP bandits and RL in interactive sequential learning.

Robustness and Safety: The second thrust addresses the critical need for robustness in interactive learning, where algorithms must make decisions potentially with misleading or corrupted data inputs. This thrust will explore the interplay between privacy and robustness, investigating how privacy protection can impact the robustness of the systems. By developing new theories and methods, this thrust will pave the way for more resilient ML systems for real-world deployment.

Fairness: The final thrust is dedicated to establishing a coherent, inclusive framework for fairness in interactive decision-making. Leveraging insights from game theory and regret minimization, this research will explore how different notions of fairness, from individual to group fairness, can be reconciled and operationalized in practical, algorithmic solutions. This effort is crucial for ensuring that the next generation of ML systems promotes equity and accessibility, offering fair and unbiased services to all users.

Applications: RL/bandits for LLM alignment and LLM as an agent for in-context RL.

Broader Impacts:

The societal, economic, and educational ramifications of this research are profound. In alignment with national priorities, such as those articulated in recent executive orders on AI ethics, this project aims to make significant contributions to the domains of AI privacy and safety, fostering a more secure, equitable, and trustworthy technological landscape. Economically, the enhancement of personalized services grounded in trustworthiness principles stands to benefit service providers and consumers alike, catalyzing growth and innovation across industries. Educationally, the project will serve as a beacon for pioneering educational initiatives, developing comprehensive courses on trustworthy decision-making. These courses will not only attract a diverse body of students to the cutting edge of machine learning but also engage younger audiences through outreach, inspiring the next generation at Wayne State University and beyond.