AI Safety: From Reinforcement Learning to Foundation Models

Ezgi Korkmaz

From learning to make sequential decisions from raw high-dimensional data to interacting with humans solely based on learning a model of probability distributions over tokens, i.e. large language models, the machine learning field is experiencing immense progress towards achieving intelligent agents making important decisions for humanity in everyday life. The advancements of reinforcement learning further fuel the research on foundation models aiming to build large language agents that can reason, and are responsible, aligned, unbiased and robust. While these models are currently being deployed in high stake decision making with societal impact, the concerns on the reliability, robustness and safety of these models remains to be an open problem.

This tutorial will introduce a principled analysis of current learning paradigms on responsible, robust and safe machine learning, and further will reveal how and why the current learning paradigms fall short on providing safety, robustness and generalization.

Website: https://aaai.org/conference/aaai/aaai-25/tutorial-and-lab-list/#TQ10

References:
Ezgi Korkmaz. Understanding and Diagnosing Deep Reinforcement Learning. International Conference on Machine Learning, ICML 2024.
Ezgi Korkmaz. Adversarial Robust Deep Reinforcement Learning Requires Redefining Robustness. AAAI Conference on Artificial Intelligence, AAAI 2023.
Ezgi Korkmaz et al. Detecting Adversarial Directions in Deep Reinforcement Learning to Make Robust Decisions. International Conference on Machine Learning, ICML 2023.
Ezgi Korkmaz. Deep Reinforcement Learning Policies Learn Shared Adversarial Features Across MDPs. AAAI Conference on Artificial Intelligence, AAAI 2022.
Ezgi Korkmaz. Investigating Vulnerabilities of Deep Neural Policies. Conference on Uncertainty in Artificial Intelligence (UAI), Proceedings of Machine Learning Research (PMLR), PMLR 2021.