Safe Behavior in Open Worlds: AI Methods for Learning and Acting in the Presence of Unknown Unknowns

Abstract

The development of AI technology has progressed from working with "known knowns"—AI planning and problem solving in deterministic, closed worlds—to working with "known unknowns"—planning and learning in uncertain environments based on probabilistic models of those environments. A critical challenge for future AI systems is to behave safely and conservatively in open worlds, where most aspects of the environment are not modeled by the AI agent—the "unknown unknowns". Our team, with deep experience in machine learning, probabilistic modeling, and planning, will develop principles, evaluation methodologies, and algorithms for learning and acting safely in the presence of the unknown unknowns. For supervised learning, we will develop UU-conformal prediction algorithms that extend conformal prediction to incorporate nonconformity scores based on robust anomaly detection algorithms. This will enable supervised learners to behave safely in the presence of novel classes and arbitrary changes in the input distribution. For reinforcement learning, we will develop UU-sensitive algorithms that act to minimize risk due to unknown unknowns. A key principle is that AI systems must broaden the set of variables that they consider to include as many variables as possible in order to detect anomalous data points and unknown side-effects of actions.