Explainability for Safety in Reinforcement Learning

Authors: Tongtong Liu, Md Asifur Rahman, Joe McCalmon, Sarra Alqahtani

Summary:

In recent years, interpretations of complex AI systems' behavior have become a crucial field of study. Although prior studies have made significant advancements in interpreting a wide range of machine learning algorithms, there is a great demand for relevant studies in the field of reinforcement learning (RL) and explaining agents' behavior. Also, the lack of explainability implies that the optimal strategies of agents cannot be used to improve our understanding of their safety. Explaining the safety measures of an RL agent is important in ensuring its effectiveness. In this work, we improved our explainable RL method, CAPS, and use it for the comparative analysis of the safety behavior pattern of our proposed safety algorithm against the existing RL baseline. We hypothesize that RL safety will result if CAPS transparently offers enough explanations of agents' decisions and actions in terms of safety.

ACM Author Affiliations: Tongtong Liu: Wake Forest University; Md Asifur Rahman: Wake Forest University; Joe McCalmon: Wake Forest University; Sarra Alqahtani: Wake Forest University