Trustworthy in Reinforcement Learning

This workshop seeks to motivate researchers in the realm of reinforcement learning systems to integrate safety considerations into both the creation and utilization of these systems. The term 'safety' is defined extensively, covering the prevention of self-injury, environmental damage, and significant societal costs.

Workshop Topics:

Emphasizing this broad understanding of safety, we are calling for extended abstracts on topics such as:

- Developing safe RL policies
- Verifying the robustness of RL policies
- Safe RLHF in LLMs
- Safety issues in control systems
- Practical applications, demonstrations, or defining problems
- Simulation tools and datasets

Workshop Organizers:

Ronghui Mu, a lecturer at the University of Exeter, has contributed several works on RL safety to the AAAI conference and has around five years of experience in the field of robust Deep Neural Networks (DNNs).

Gaojie Jin, University of Exeter.

Important Dates:

- Workshop Paper Submission Deadline: September 15, 2024
- Notification of Acceptance: October 15, 2024

Papers are to be submitted as PDF via the site: https://edas.info/N32633.
Please select the corresponding workshop when submitting your paper.