Addressing Goal Misgeneralization in Offline Reinforcement Learning with Natural Language

Giulio Starace

28th February, 2023

Research Question

Testing citations Abbeel and Ng (2004)

Literature Review

Offline Reinforcement Learning

Goal Misgeneralization

Natural Language Models

Natural Language and Reinforcement Learning

Methodology

Planning

References

Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA. Association for Computing Machinery.
hello