Our target tasks are cool and all, but here's what I'm really interested in solving!

Thorben Klamt - 10.04.2024

Advanced Topics in Reinforcement Learning, , Gottfried Wilhelm Leibniz University Hannover

Security, safety, interpretability and robustness of RL policies

- Very high potential of RL over broad range of real world applications
- High expenses in case of failure in broad range of applications as well

Security, safety, interpretability and robustness of RL policies

- Very high potential of RL over broad range of real world applications
- High expenses in case of failure in broad range of applications as well
- Need for mechanisms to increase security and reliability
  - Interpretability within Deep Reinforcement Learning
    - Interpretable model extraction
    - Sub-policy joining and unsupervised classification
    - Deep layer interpretable model extraction, sparsity etc.
  - Ensemble agreement, OOD detection etc.