Towards Robust End-to-End Alignment

Lê Nguyên Hoàng
EPFL
Safe AI 2019
Alice

Reinforcement Q/policy-learning Interruptibility Lookahead Safe exploration
Dave
World Model
Feature learning
Bayes
MCMC
World description
Charlie
Desirability
Inverse RL
CEV
Social choice
Trust
Bob
Incentive
Reward hacking
Wireheading
Interactive proof
Backpropagation
Transdisciplinary problems
Decentralization, Byzantine resilience, heuristics, specialization, secure messaging, burden assignment...