***Introduction***

Disadvantages of other methods:

Q-learning: fails on many simple problems and is poorly understood

Policy Gradient: poor data efficiency and robustness

TRPO: complicated and is not compatible with architectures that include noise (dropout) or parameter sharing (between the policy and value function, or with auxiliary tasks)

Goal of PPO: Attains the data efficiency and reliable performance of TRPO