PPO Algorithm

1: **Input:** initial policy parameters θ_0, clipping threshold ϵ
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Collect set of partial trajectories D_k using policy $\pi_k = \pi(\theta_k)$
4: Estimate advantages $\hat{A}_t^{\pi_k}$ using any advantage estimation algorithm
5: Compute policy update:

$$\theta_{k+1} = \arg\max_{\theta} L_{\theta_k}^{\text{CLIP}}(\theta)$$

6: Perform K steps of minibatch SGD (via Adam), where:

$$L_{\theta_k}^{\text{CLIP}}(\theta) = E_{\tau \sim \pi_k}\left[\sum_{t=0}^{T} \min\left(r_t(\theta)\hat{A}_t^{\pi_k}, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t^{\pi_k}\right)\right]$$

7: **end for**