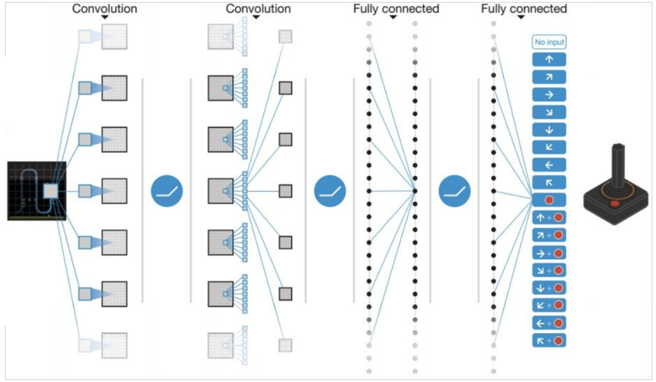
NEURAL INFORMATION PROCESSING SYSTEMS

Policy Learning Using Weak Supervision

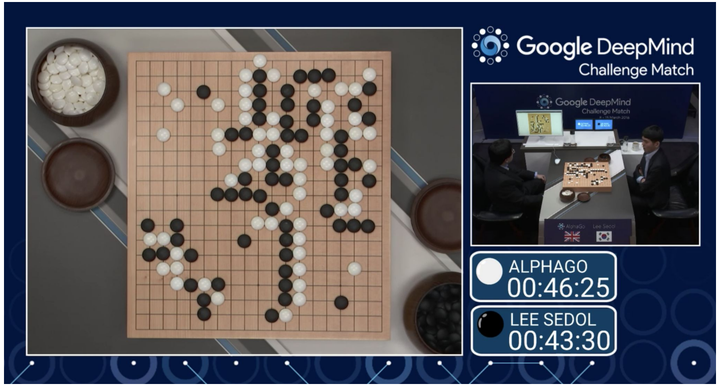
Jingkang Wang*[1,2], Hongyi Guo*[3], Zhaowei Zhu*[4], Yang Liu[4]

UNIVERSITY OF TORONTO
VECTOR INSTITUTE
Northwestern University
UNIVERSITY OF CALIFORNIA SANTA CRUZ

Deep Learning in Sequential Decision Making

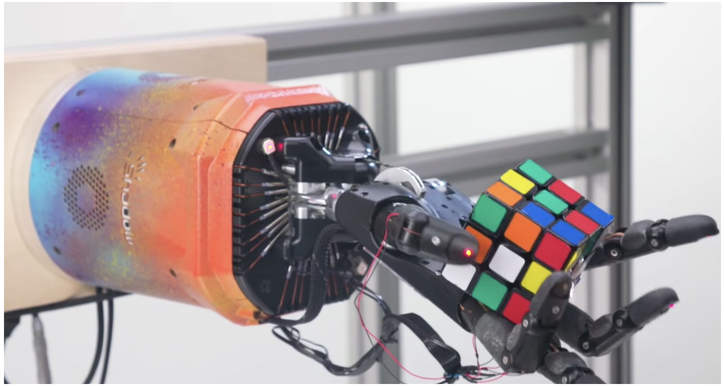


Atari2600 Games [Mnih et al., 2015]



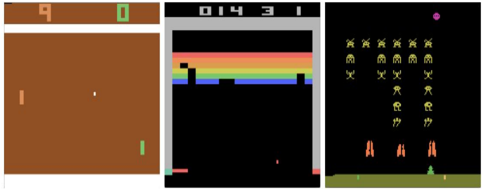
AlphaGo [Silver et al., 2017]

Self-Driving [Amini et al., 2020]

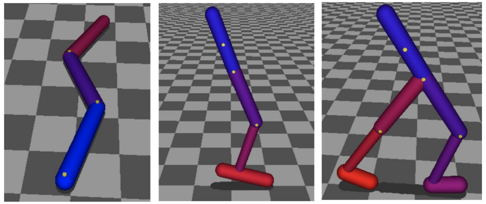


Self-Driving [OpenAI, 2019]

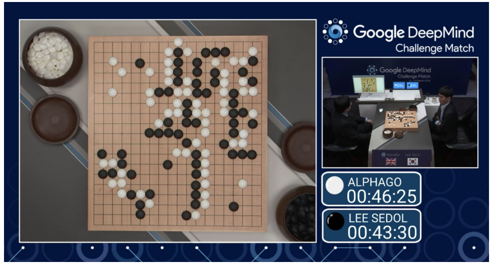
Markov Decision Process (MDP)



[Mnih et al., 2013]



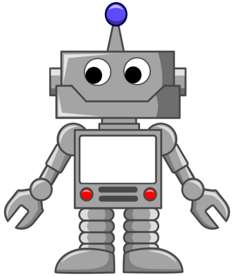
[Schulman et al., 2015]



[AlphaGo versus Lee Sedol]

MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$

Action $a_t \in \mathcal{A}$

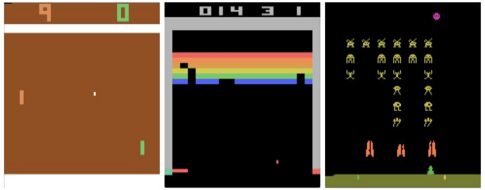


Agent

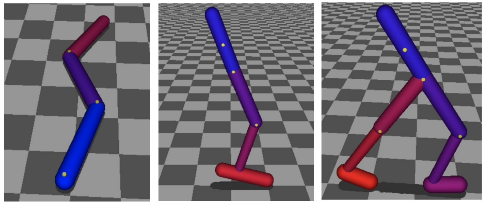
$\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$

Environment

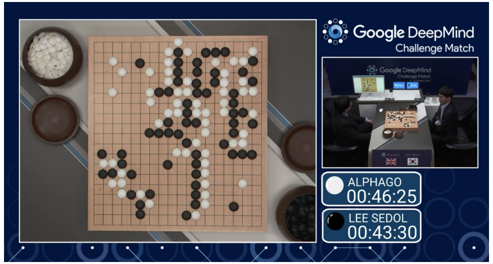
Markov Decision Process (MDP)



[Mnih et al., 2013]



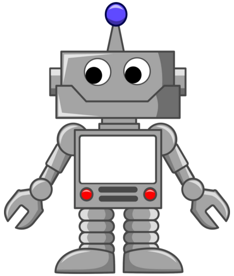
[Schulman et al., 2015]



[AlphaGo versus Lee Sedol]

MDP: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$

Action $a_t \in \mathcal{A}$



Agent

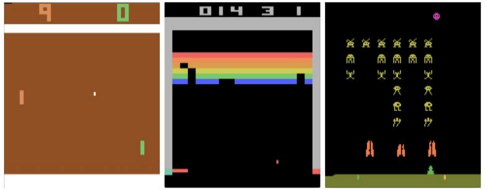
$\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$

Environment

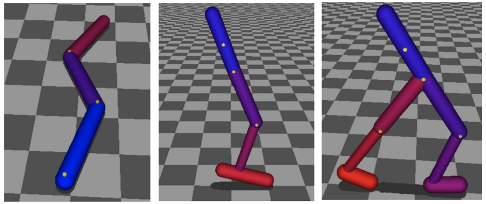
State $s_{t+1} \in \mathcal{S}$

Reward $r(s_t, a_t) \in \mathcal{R}$

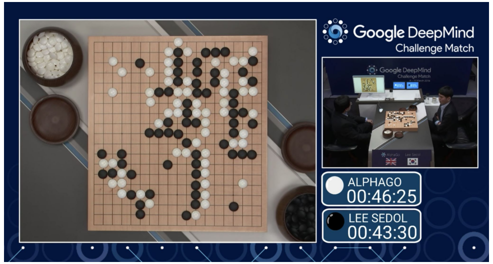
Markov Decision Process (MDP)



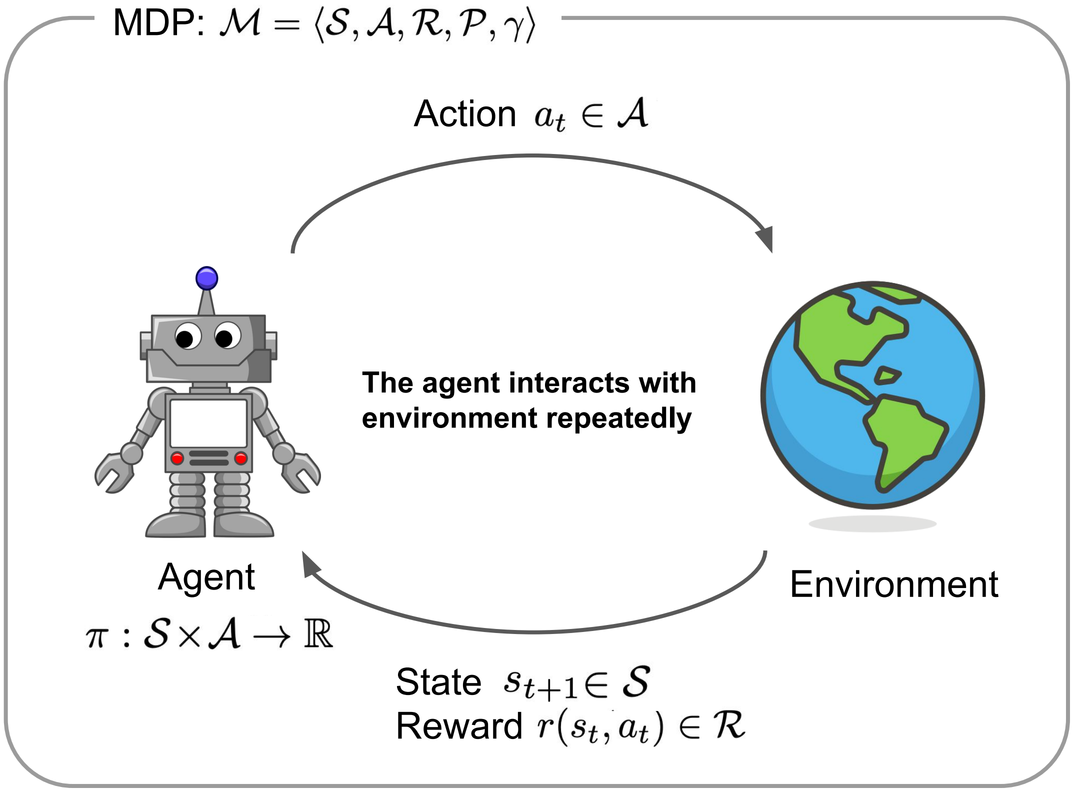
[Mnih et al., 2013]



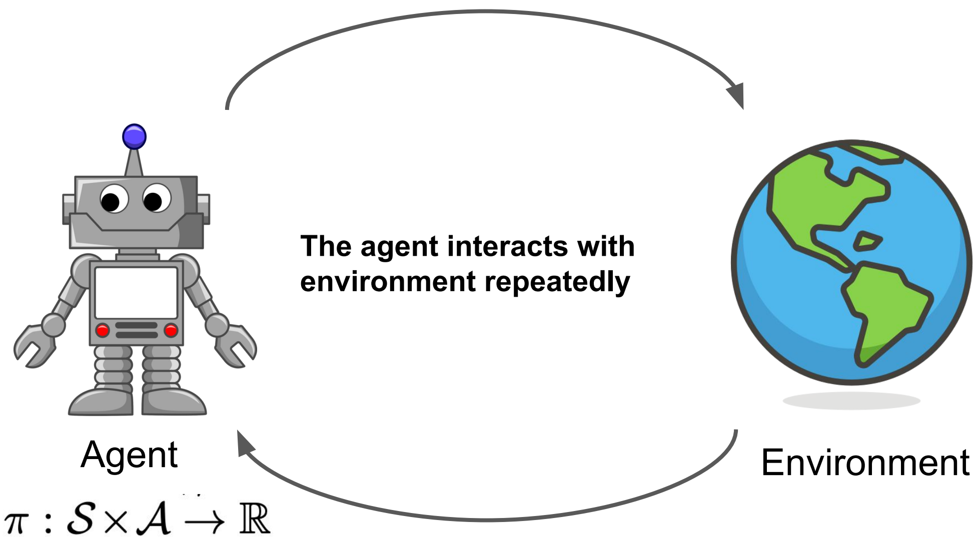
[Schulman et al., 2015]



[AlphaGo versus Lee Sedol]



Reinforcement Learning



Generated trajectory:

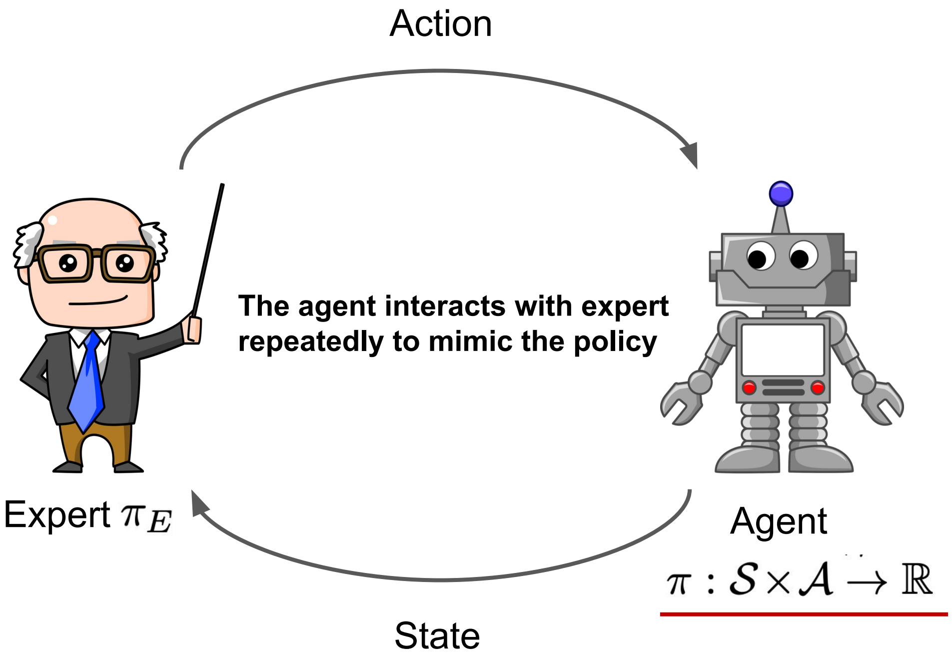
$$\tau = \{(s_t, a_t, r_t)\}_{t=0}^{T}$$

Objective: maximize the expected reward

$$J(\pi) = \mathbb{E}_{(s_t, a_t, r_t) \sim \tau}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

Supervision signal: *reward*

Imitation Learning



Expert demonstrations

$$D_E = \{(s_i, a_i)\}_{i=1}^{N}$$

Behavioral Cloning Objective:
maximize the log-likelihood

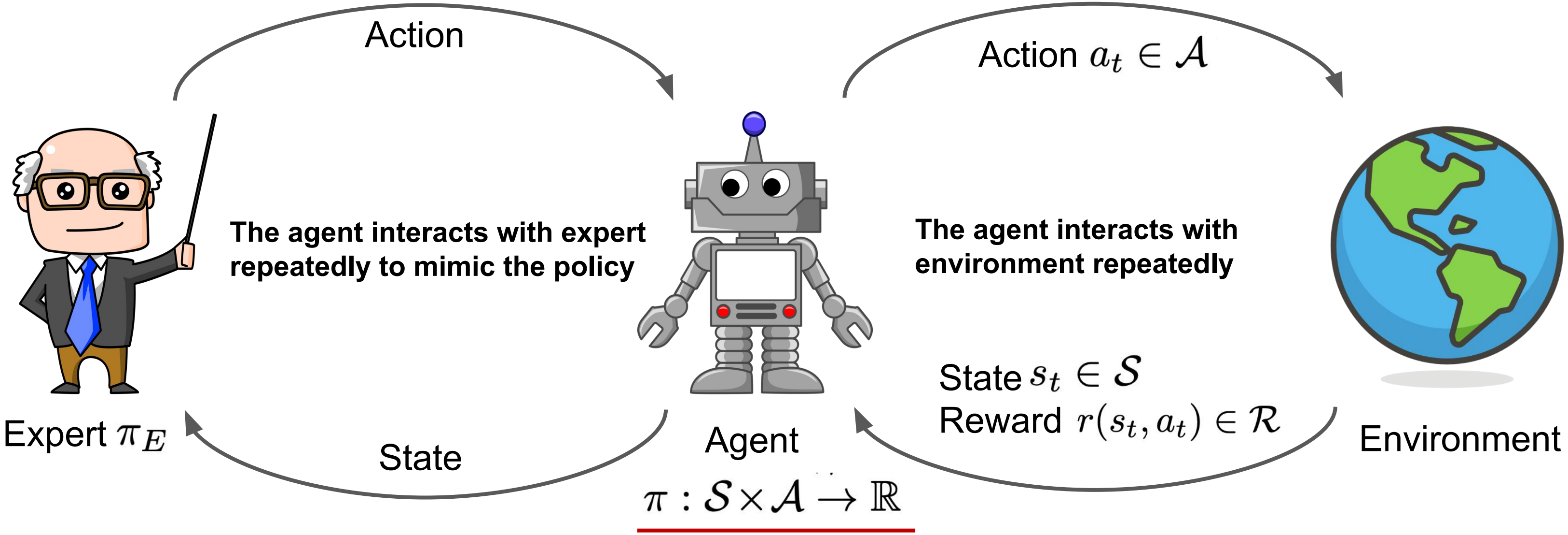
$$J(\pi) = \mathbb{E}_{(s,a)\sim\mathcal{D}_E}[\log \pi(a|s)]$$

Supervision signal: *expert action*

Hybrid Learning (RL + IL)

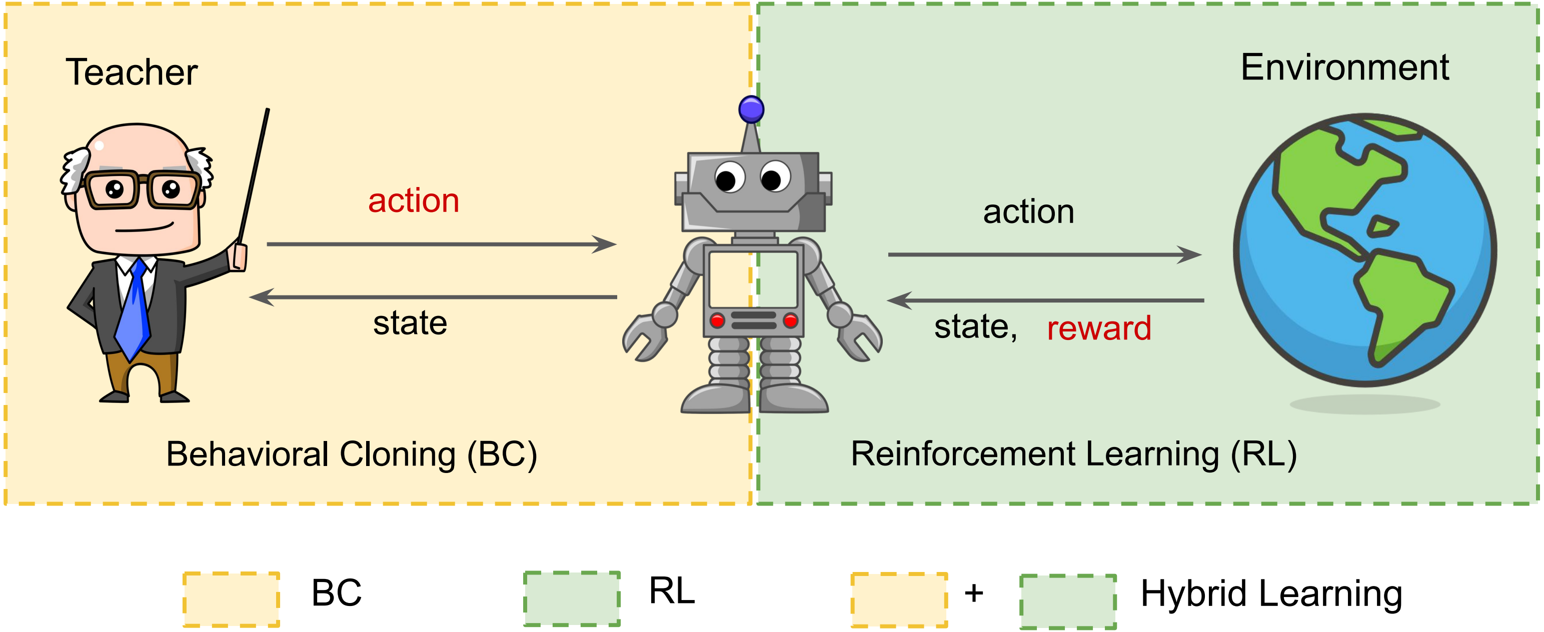
Supervision signal: *reward* + *expert action*

Hybrid objective: $J(\pi) = \lambda_1 \mathbb{E}_{(s_t, a_t, r_t) \sim \tau} \left[\sum_{t=0}^{T} \gamma^t r_t \right] + \lambda_2 \mathbb{E}_{(s,a) \sim \mathcal{D}_E} [\log \pi(a|s)]$



Policy Learning

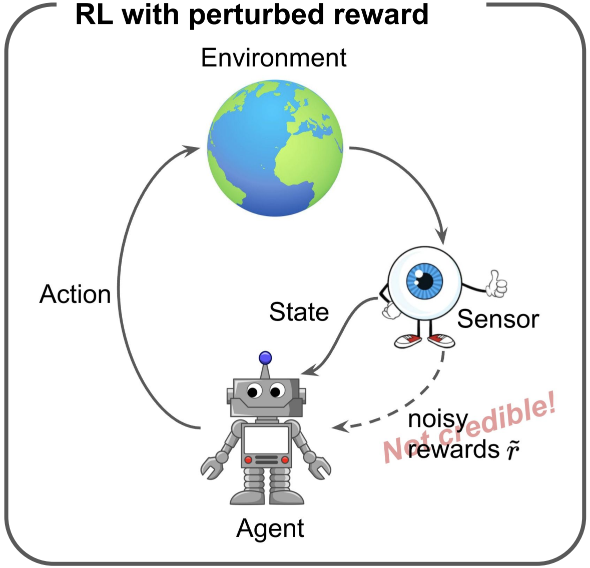
Summary:

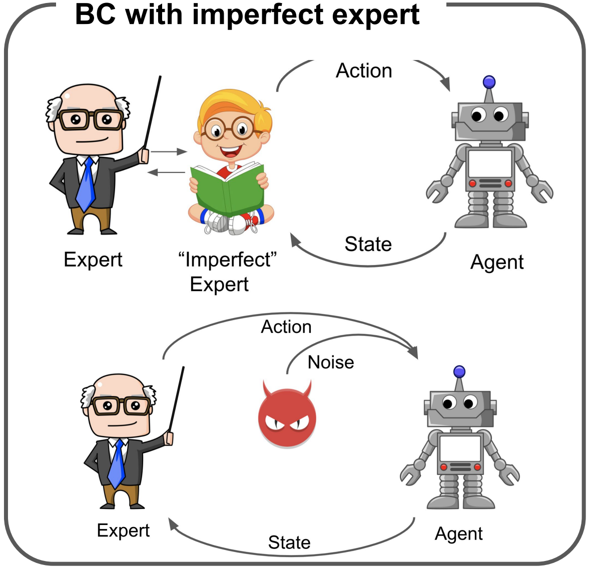


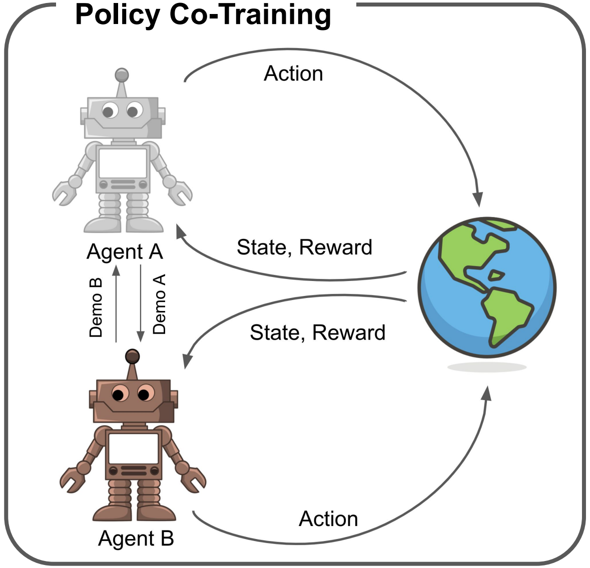
Weak supervision signals are everywhere!

Weak Supervision:

- **RL:** The reward may be collected through sensors thus noisy
- **IL:** The demonstrations by an expert are often imperfect due to limited resources

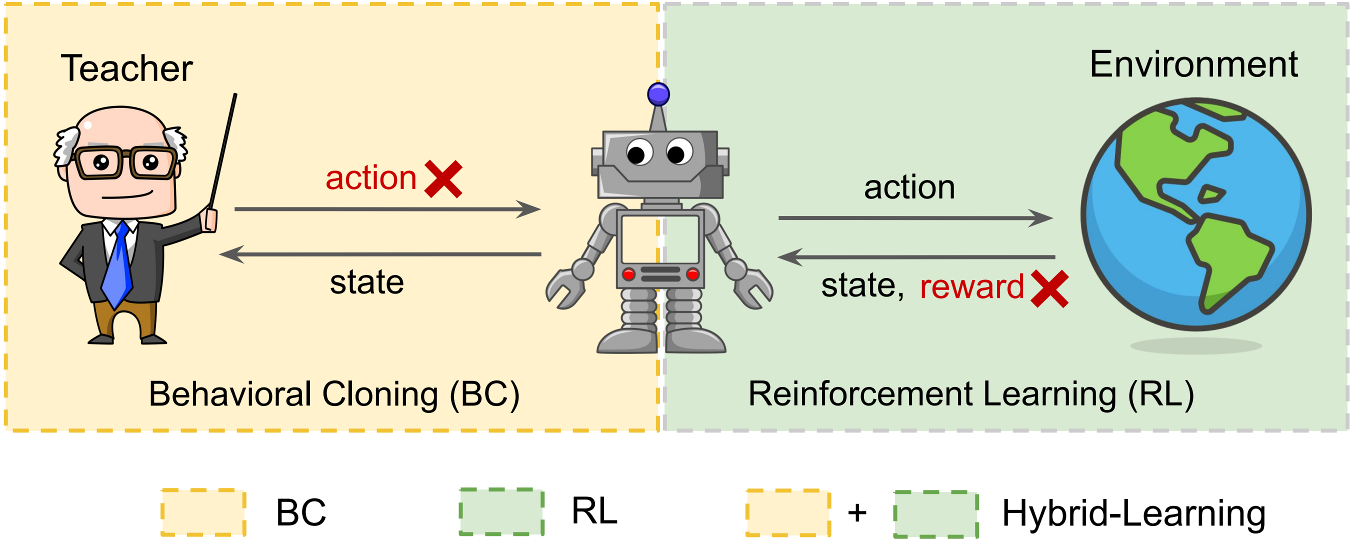






Weakly Supervised Policy Learning

Problem: Supervision signals $\tilde{Y}$ (either reward or expert's demonstrations) are ***not credible!***

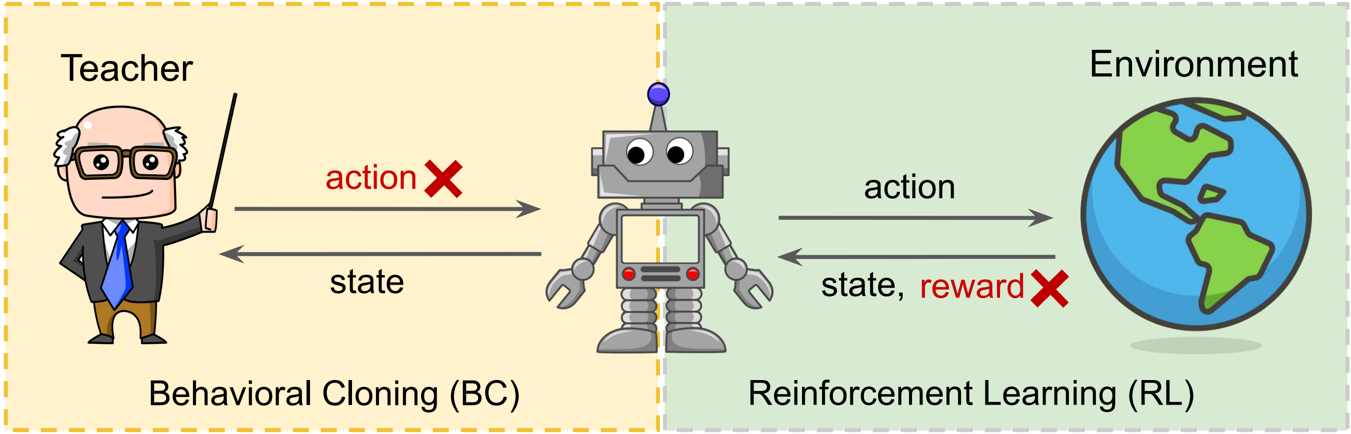


Weak Supervision:

- **RL:** The reward may be collected through sensors thus noisy
- **IL:** The demonstrations by an expert are often imperfect due to limited resources

Weakly Supervised Policy Learning

Weakly Supervised Policy Learning $\{(s_i, a_i), \widetilde{Y}_i\}_{i=1}^{N}$

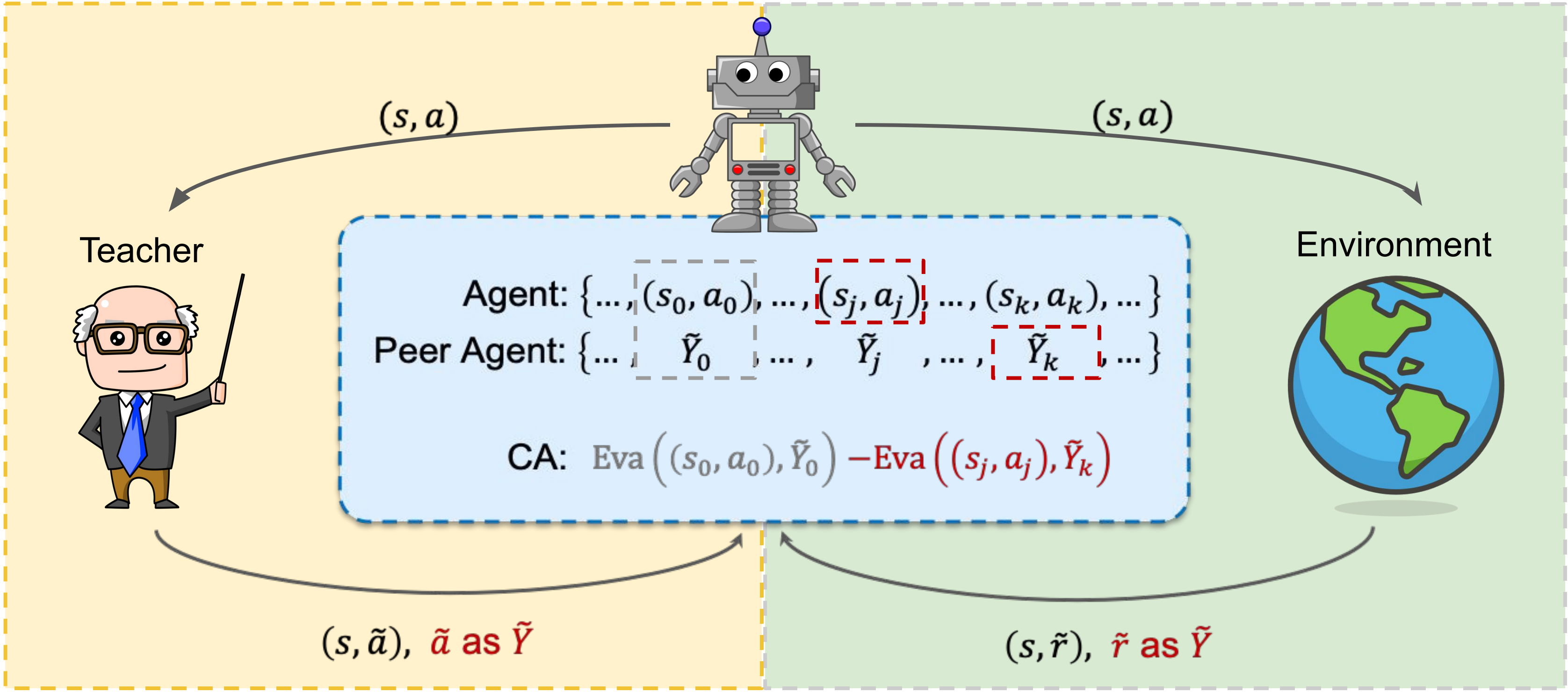


BC RL + Hybrid-Learning

- Objective: $J(\pi) = \mathbb{E}_{(s,a)\sim\tau}\left[\mathtt{Eva}_{\pi}\left((s,a),\tilde{Y}\right)\right]$

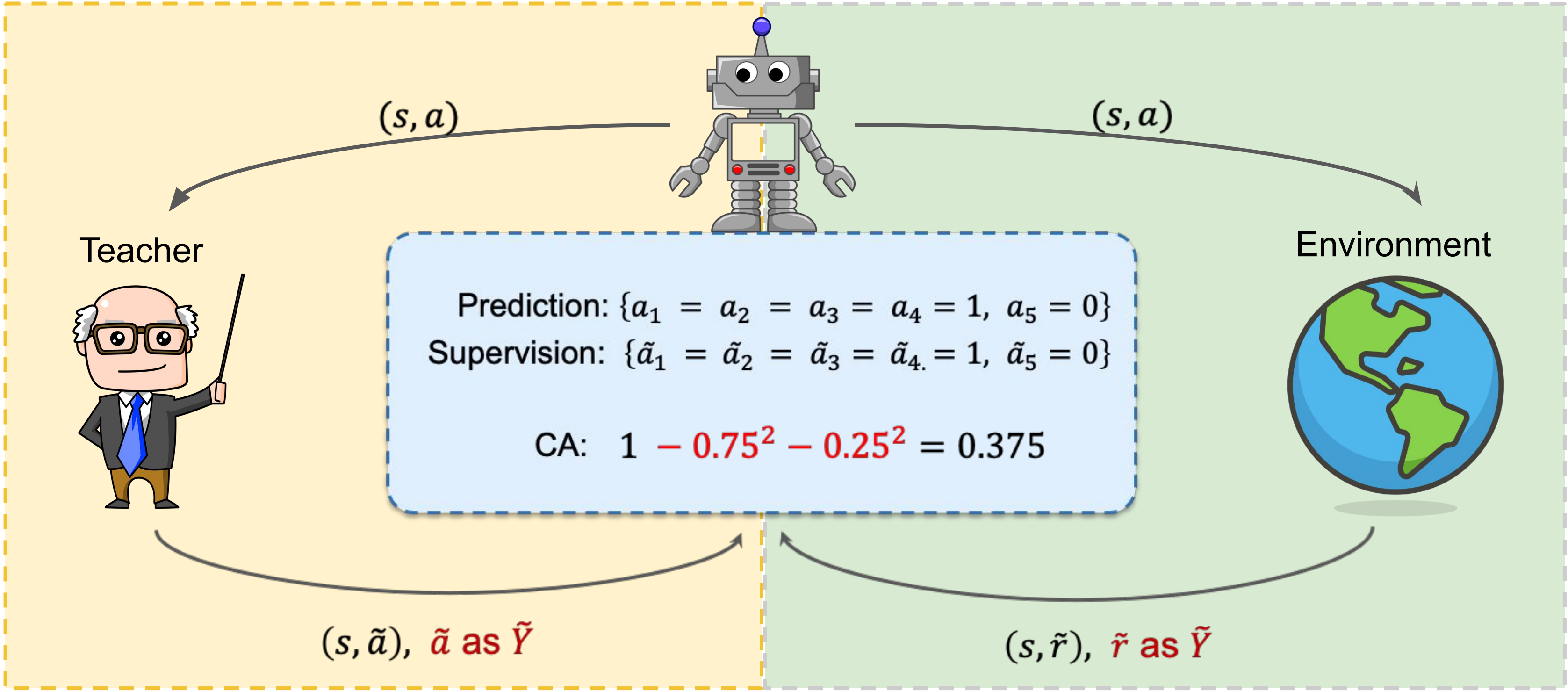
Correlated Agreement (CA)

Solution - CA with weak supervision: $\mathtt{Eva}_\pi\left((s_i, a_i), \widetilde{Y}_i\right) - \mathtt{Eva}_\pi\left((s_j, a_j), \widetilde{Y}_k\right)$



Correlated Agreement (CA)

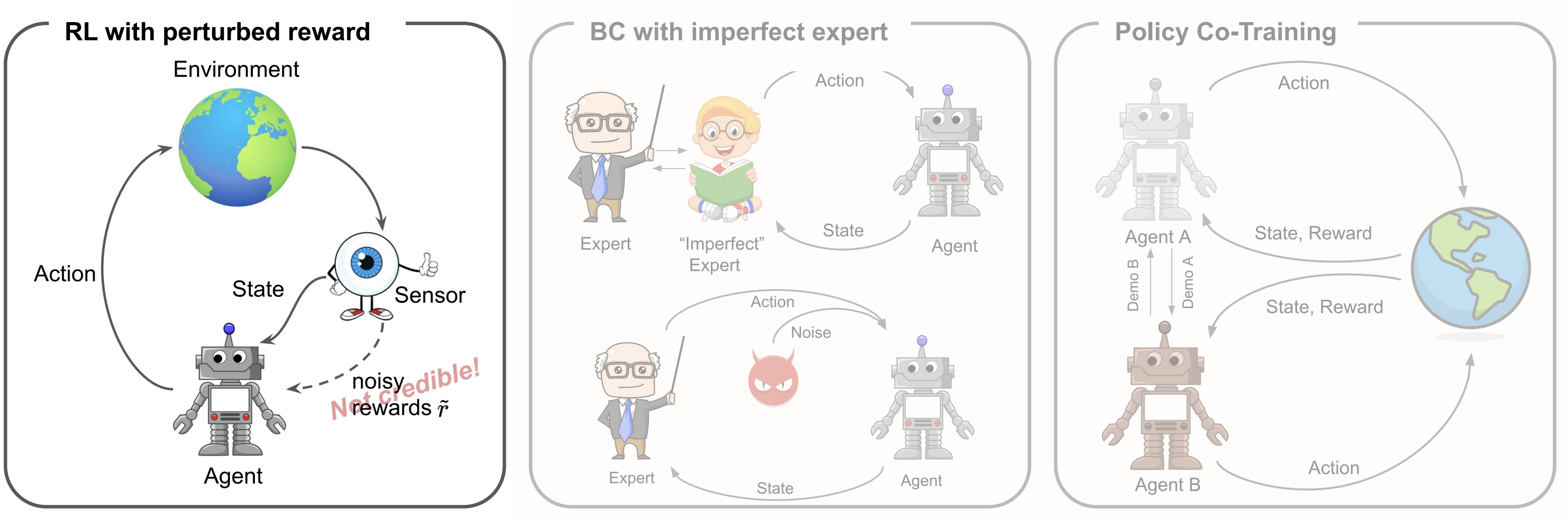
Solution - CA with weak supervision: $\texttt{Eva}_\pi\big((s_i, a_i), \widetilde{Y}_i\big) - \texttt{Eva}_\pi\big((s_j, a_j), \widetilde{Y}_k\big)$



PeerPL: A Unified Framework for Weakly Supervised PL

Solution - CA with weak supervision: $\mathtt{Eva}_\pi\left((s_i, a_i), \widetilde{Y}_i\right) - \mathtt{Eva}_\pi\left((s_j, a_j), \widetilde{Y}_k\right)$

PeerRL



PeerRL

We assume the noisy reward $\tilde{r}$ is generated following a certain function $F : \mathcal{R} \rightarrow \tilde{\mathcal{R}}$.

- Discrete with $|\mathcal{R}|$ levels.
- Characterized via an ***unknown*** matrix $\mathbf{C}^{\mathrm{RL}}_{|\mathcal{R}|\times|\mathcal{R}|}$

$$r(s,a) \xrightarrow{\mathbf{C}^{\mathrm{RL}}_{|\mathcal{R}|\times|\mathcal{R}|}} \tilde{r}(s,a)$$

PeerRL handles the noisy reward by defining the peer RL reward:

$$\tilde{r}_{\text{peer}}(s,a) = \tilde{r}(s,a) - \xi \cdot \tilde{r}'$$

where $\tilde{r}' \overset{\pi_{\text{sample}}}{\sim} \{\tilde{r}(s,a) | s \in \mathcal{S}, a \in \mathcal{A}\}$ is a reward sampled over all state-action pairs according to a fixed policy π_{sample}.

$$\tilde{r}(s,a) \xrightarrow{\mathrm{CA}} \tilde{r}_{\text{peer}}(s,a)$$

Our theory shows that peer RL rewards are robust to noisy rewards *(see Lemma 1 and Theorem 1)*.

Why Peer Reward Works?

- **Hypothesis 1:** PeerRL reduces the bias (while with larger variance like Wang et al., 2020)

noisy reward: $$\mathbb{E}[\tilde{r}] = \eta \cdot \left(\mathbb{E}[r] + \underline{\frac{e_+}{1 - e_- - e_+}} r_- + \underline{\frac{e_-}{1 - e_- - e_+}} r_+\right)$$

peer reward: $$\mathbb{E}[\tilde{r}_{\text{peer}}] = \eta \cdot (\mathbb{E}[r] - (1 - p_{\text{peer}})r_- - p_{\text{peer}} r_+)$$

potentially much larger than $(1 - p_{\text{peer}})$ and p_{peer} in high noise regime!

- **Hypothesis 2:** PeerRL helps break ties
  - "tie" states indicate that the rewards for different states are the same - unstable and uncertain
  - randomness in discretization model thus breaking ties - more informative for optimization

2-state Markov process (no actions)

$$r_1 \sim \text{clamp}[\mathcal{N}(0.6, 1), \min = 0, \max = 1]$$

$$r_2 \sim \text{clamp}[\mathcal{N}(0.4, 1), \min = 0, \max = 1]$$

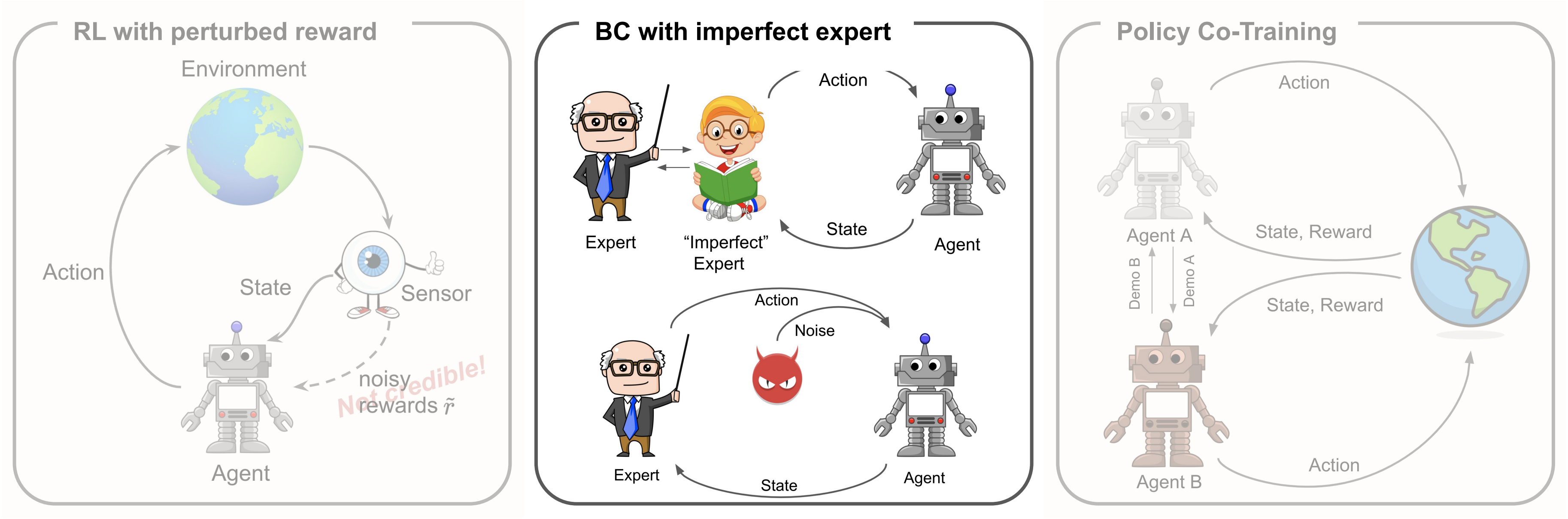
	Correct	Tie	Incorrect
Baseline	54.6%	5.6%	39.8%
PeerRL	58.0%	0.3%	41.7%

Tie breaking!

PeerPL: A Unified Framework for Weakly Supervised PL

Solution - CA with weak supervision: $\texttt{Eva}_\pi\left((s_i, a_i), \widetilde{Y}_i\right) - \texttt{Eva}_\pi\left((s_j, a_j), \widetilde{Y}_k\right)$

PeerBC



PeerBC

Available weak demonstrations $\{(s_i, \tilde{a}_i)\}_{i=1}^{N}$ where $\tilde{a}_i \sim \tilde{\pi}_E(\cdot|s_i)$

- The noisy action $\tilde{a}_i$ is independent of the state given the deterministic expert action $\pi_E(s)$
- The noise is characterized by an ***unknown*** confusion matrix $\mathbf{C}^{\mathrm{BC}}_{|\mathcal{A}|\times|\mathcal{A}|}$

$$a_i \xrightarrow{\mathbf{C}^{\mathrm{BC}}_{|\mathcal{A}|\times|\mathcal{A}|}} \tilde{a}_i$$

Again, we use CA with weak supervision to handle the noise

$$\text{Eva}^{\mathrm{BC}}_{\pi}((s_i, a_i), \tilde{a}_i) \xrightarrow{\mathrm{CA}} J^{\mathrm{BC}}(\pi_\theta)$$

- Taking cross-entropy loss for example

$$J^{\mathrm{BC}}(\pi_\theta) = \mathbb{E}\Big[\text{Eva}^{\mathrm{BC}}_{\pi}((s_i, a_i), \tilde{a}_i)\Big] - \xi \cdot \mathbb{E}\Big[\text{Eva}^{\mathrm{BC}}_{\pi}((s_j, a_j), \tilde{a}_k)\Big]$$

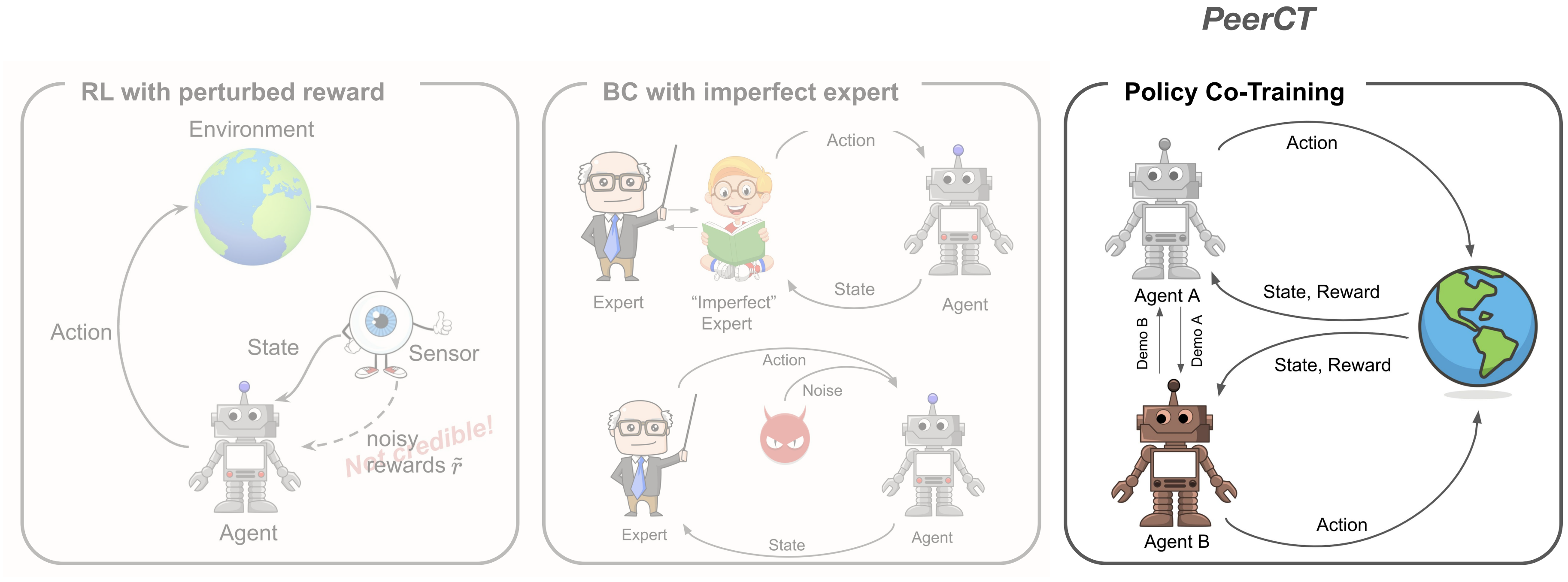
where $\text{Eva}^{\mathrm{BC}}_{\pi}(s, a), \tilde{a}) = -\ell(\pi_\theta, (s, \tilde{a})) = \log \pi_\theta(\tilde{a}|s).$

- PeerBC converges to the expert policy when observing a sufficient amount of weak demonstrations

(see Theorem 2)

PeerPL: A Unified Framework for Weakly Supervised PL

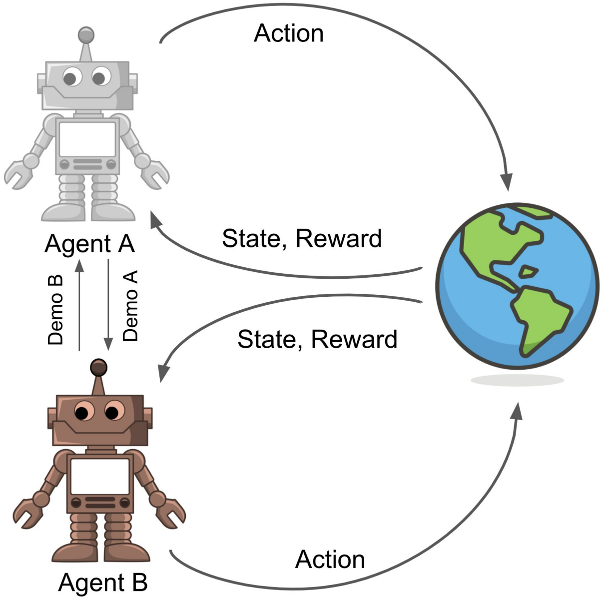
Solution - CA with weak supervision: $\texttt{Eva}_\pi\left((s_i, a_i), \widetilde{Y}_i\right) - \texttt{Eva}_\pi\left((s_j, a_j), \widetilde{Y}_k\right)$



PeerCT

Policy Co-Training (Song et al., 2019) is an instance of hybrid policy learning

- Two agents A,B with policies π^A and π^B that receive partial observations
- Agents are trained jointly to learn with rewards and noisy demonstrations from each other.
- For instance, consider agent A
    - Besides interacting with environment, A also receives $\{s_i, \pi_B(s_i)\}$ from agent B
    - We consider π^B as the noisy version of the optimal policy



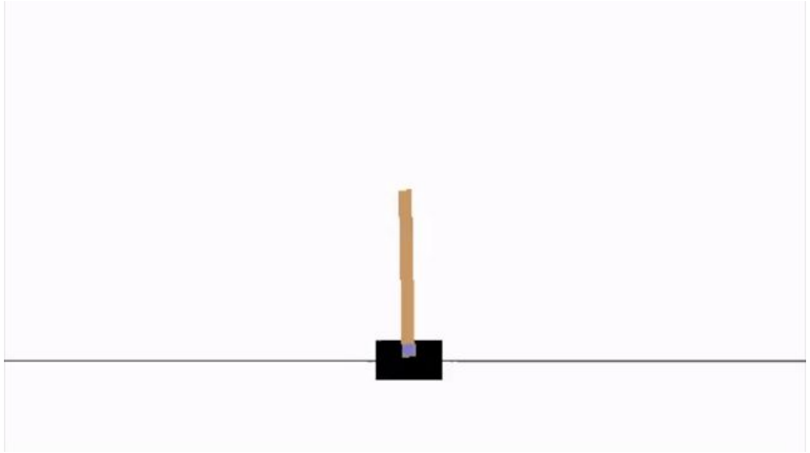
Similar to the PeerBC setting, we use CA with weak supervision to handle the noise in imperfect demonstrations

$$J^{\mathrm{CT}}(\pi_\theta) = \mathbb{E}\Big[\text{Eva}_\pi^{\mathrm{RL}}\big((s_i^A, a_i^A), r_i^A\big) + \text{Eva}_\pi^{\mathrm{BC}}\big((s_i^A, a_i^A), a_i'^B\big)\Big]$$
$$- \xi \cdot \mathbb{E}\Big[\text{Eva}_\pi^{\mathrm{BC}}\big((s_j^A, a_j^A), a_k'^B\big)\Big]$$

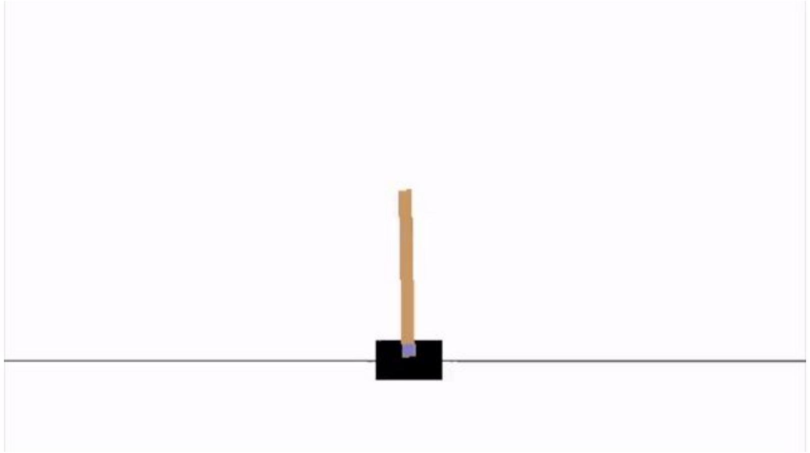
Jialin Song, Ravi Lanka, Yisong Yue, and Masahiro Ono. Co-training for policy learning. In 429 UAI, page 441. AUAI Press, 2019

An example of PeerRL on CartPole

- RL with Noisy Rewards ($e_- = e_+ = 0.4$)
- training 10,000 frames using Dueling-DQN



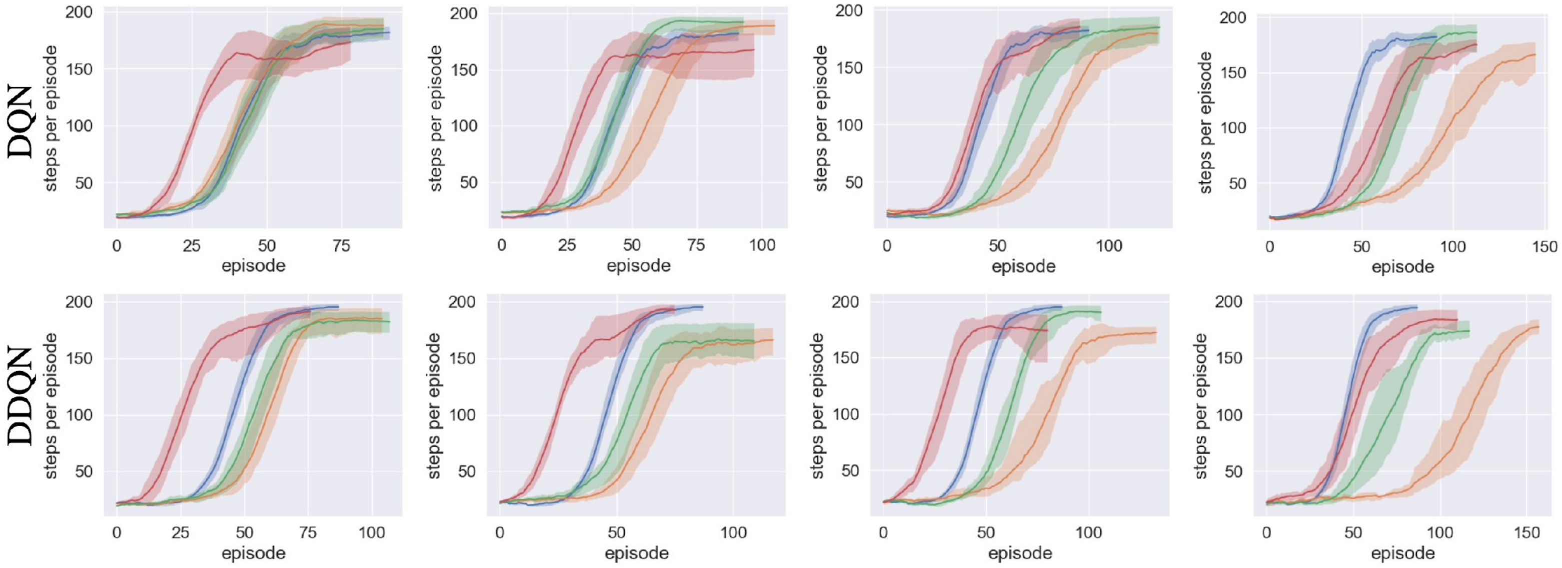
noisy reward $\tilde{r}$



peer reward $\tilde{r}_{\text{peer}}$

PeerRL recovers true reward signals

- ***CartPole***: training DDQN for 10,000 steps on, binary reward: $\{-1, 1\}$
- symmetric noise: $e = e_- = e_+$

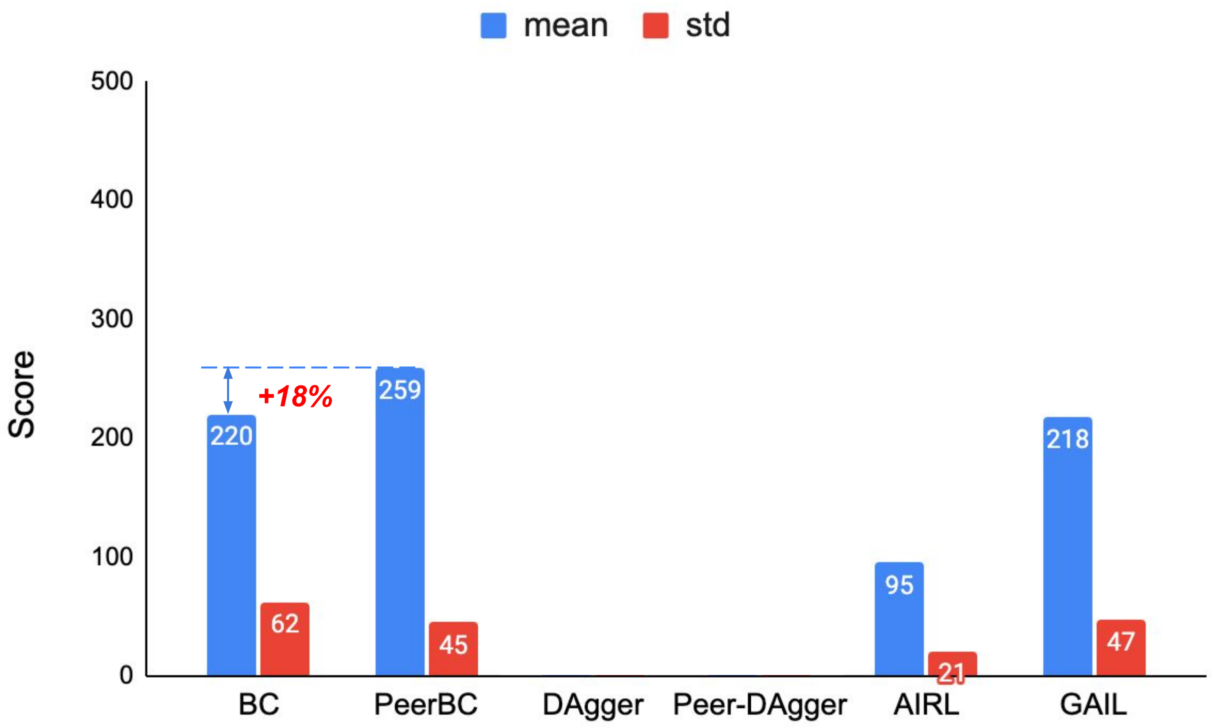


(a) $e = 0.1$ (b) $e = 0.2$ (c) $e = 0.3$ (d) $e = 0.4$

true reward | noisy reward | surrogate reward (Wang et al, 2020) | PeerRL (ours)

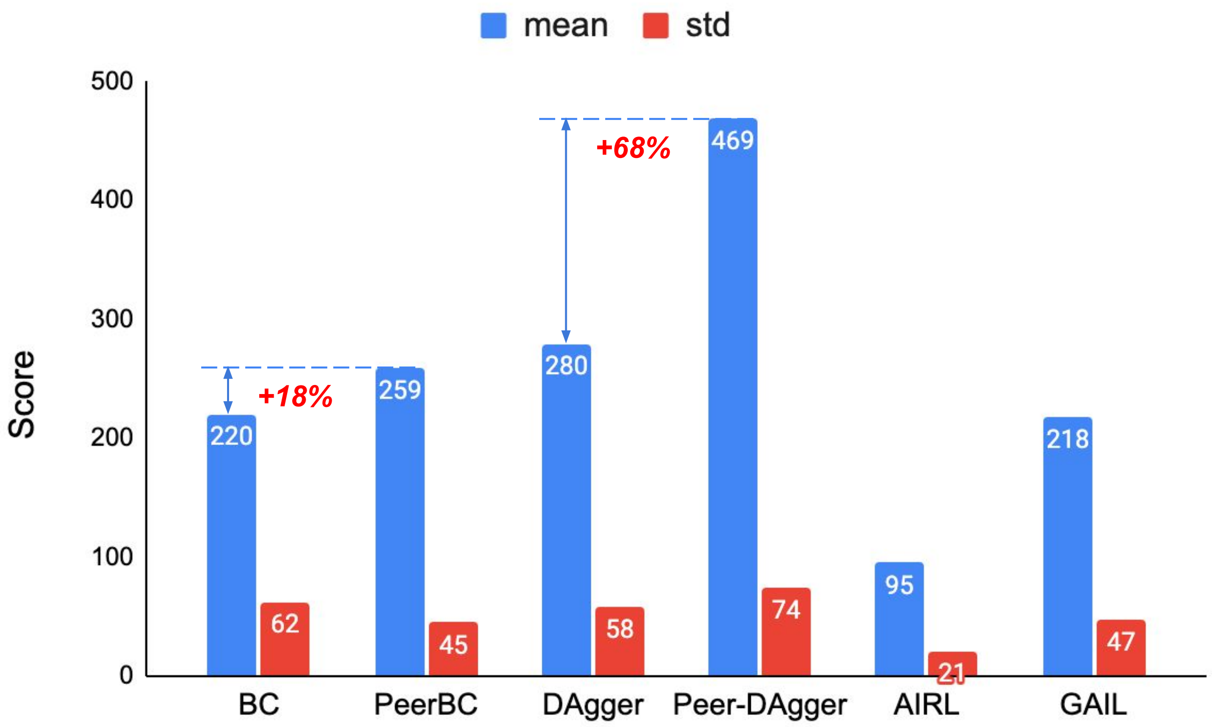
PeerBC recovers true expert signals

- ***CartPole-v1***: train an imperfect RL model with PPO algorithm, unroll 16 episodes



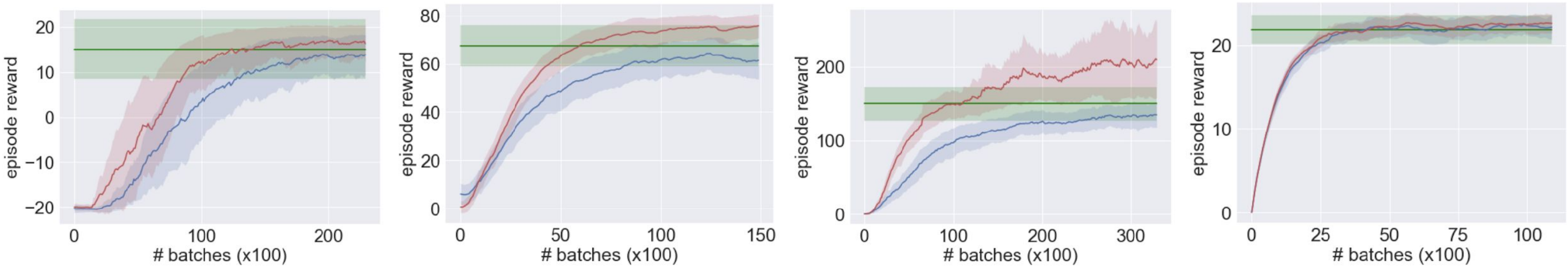
PeerBC recovers true expert supervision signals

- ***CartPole-v1***: train an imperfect RL model with PPO algorithm, unroll 16 episodes



PeerBC recovers true expert signals

- ***Atari games***: train an imperfect RL model with PPO algorithm
- weak expert = 70%~90% as good as fully converged agent
- collect demonstrations using weak expert and generate 100 trajectories for each environment
- Note that no synthetic noise is added in the experiments

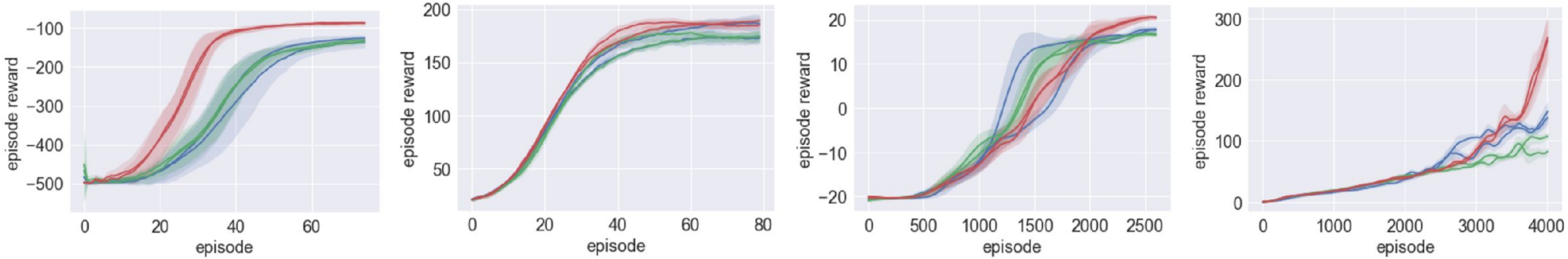


(a) Pong (b) Boxing (c) Enduro (d) Freeway

Standard BC Weak expert PeerBC (ours)

PeerCT recovers true reward signals

- ***Continuous Control/Atari***: adopt the exact same setting as Song et al., 2019 without any synthetic noise included
- removes all even index coordinates in the state vector (view-A) or removing all odd index ones (view-B)
- implies the potential of our approach to deal with natural noise in real-world applications

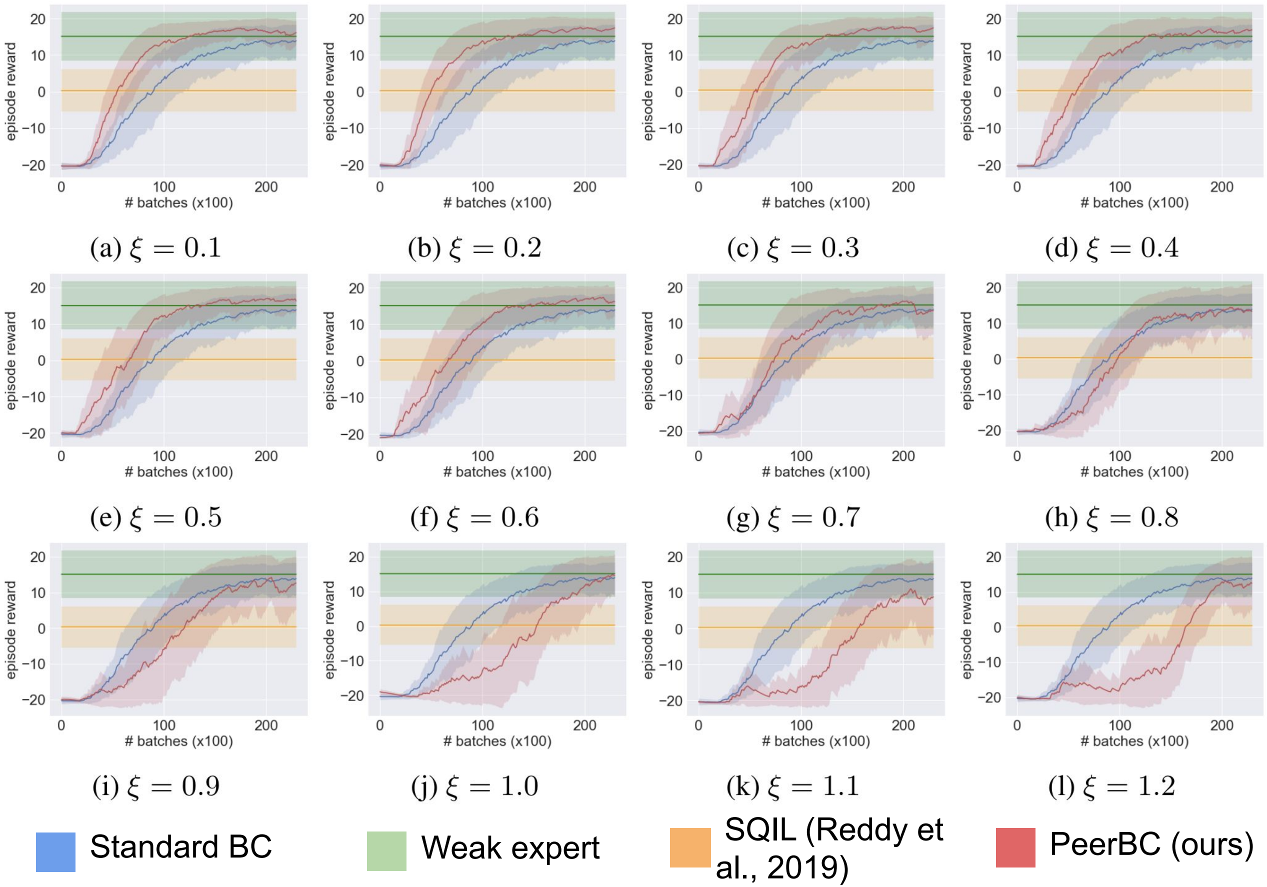


(a) Acrobot (b) CartPole (c) Pong (d) Breakout

Single view
Co-Training (Song et al., 2019)
Peer Co-Training (ours)

Sensitivity of over-agreement penalty

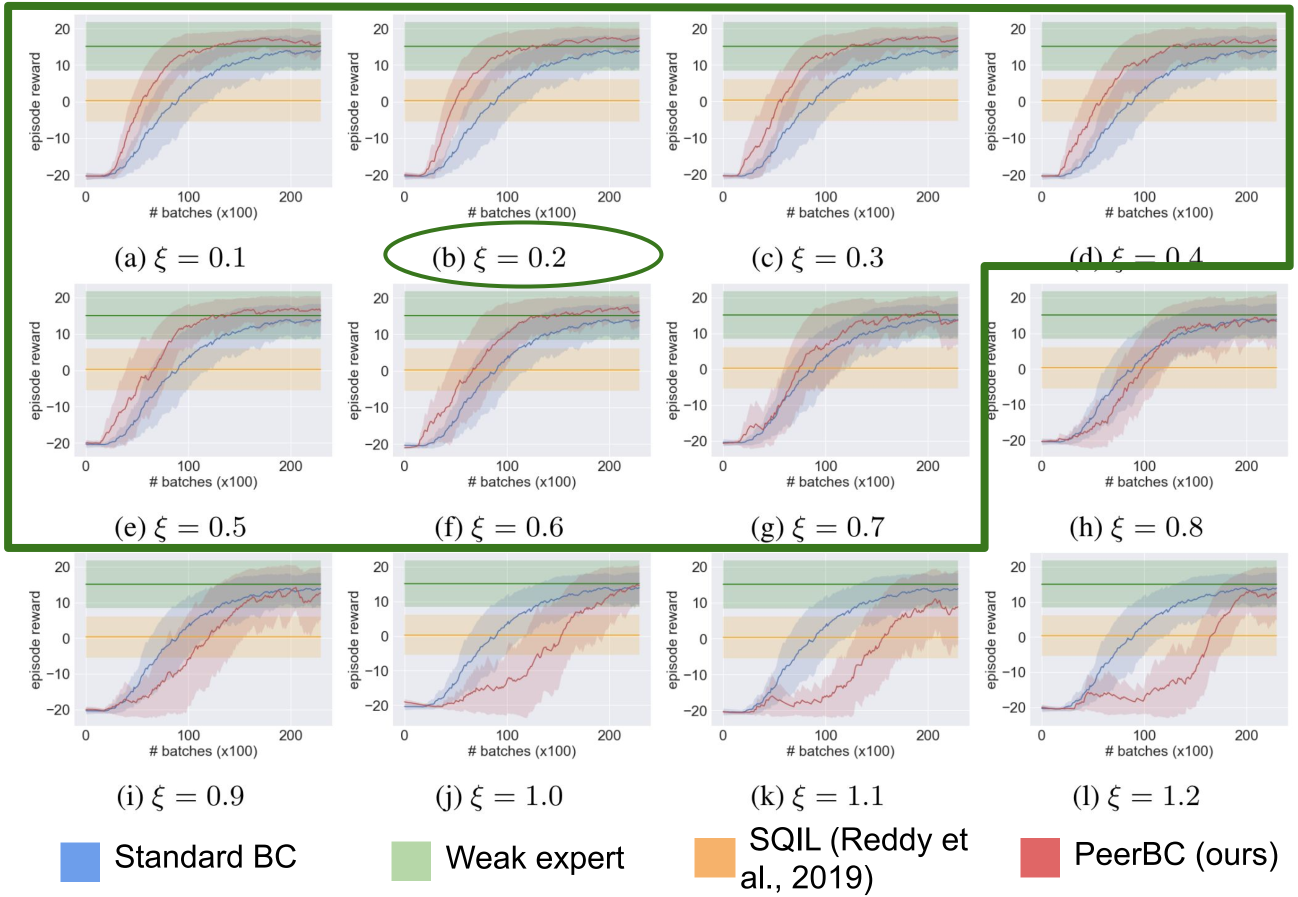
Atari - Pong



Sensitivity of over-agreement penalty

Atari - Pong

Our method works robustly in a wide range of ξ !



(a) $\xi = 0.1$ (b) $\xi = 0.2$ (c) $\xi = 0.3$ (d) $\xi = 0.4$

(e) $\xi = 0.5$ (f) $\xi = 0.6$ (g) $\xi = 0.7$ (h) $\xi = 0.8$

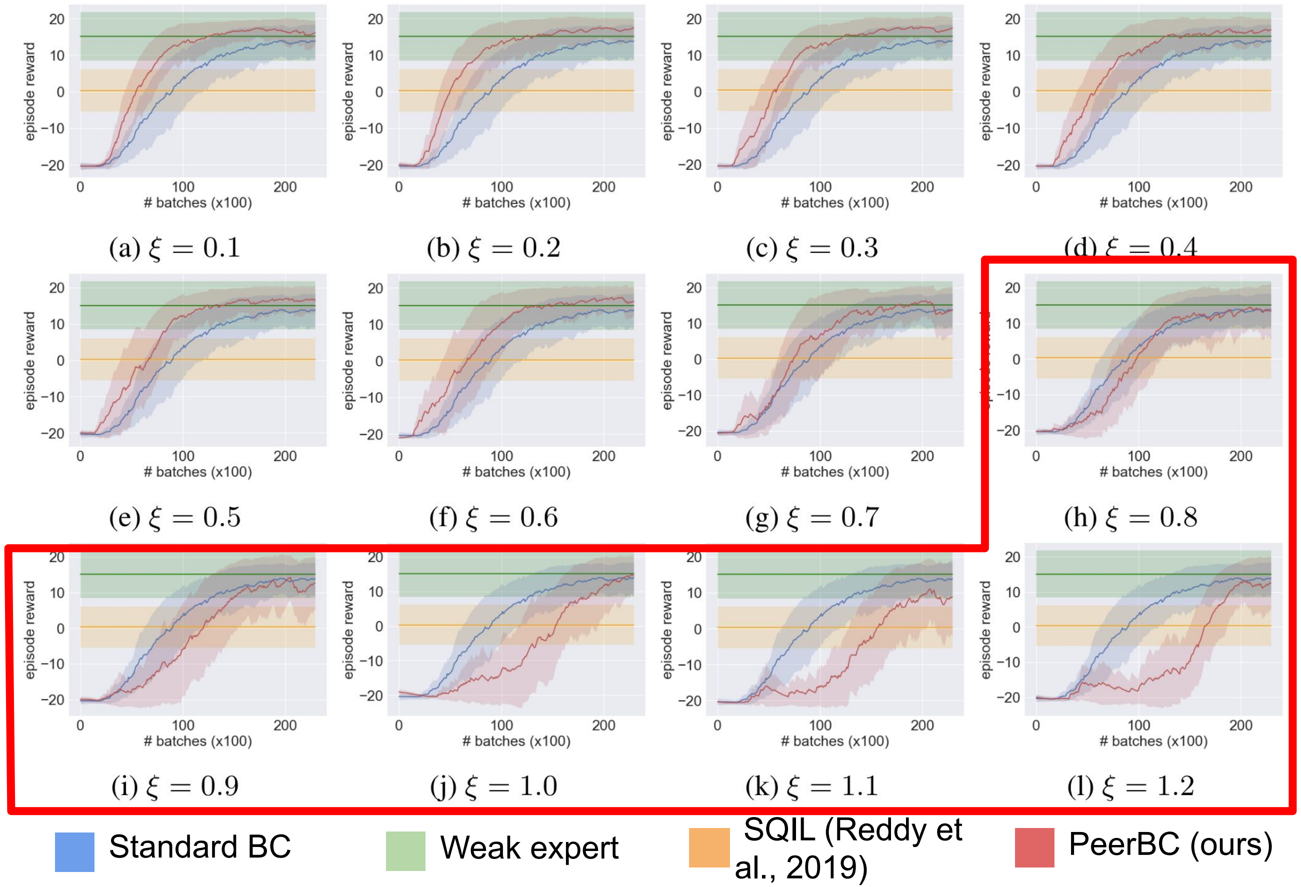
(i) $\xi = 0.9$ (j) $\xi = 1.0$ (k) $\xi = 1.1$ (l) $\xi = 1.2$

Sensitivity of over-agreement penalty

Atari - Pong

Our method works robustly in a wide range of ξ !

Overly large penalty introduces too much noise



Conclusion

- We provided a unified formulation of the ***weakly supervised policy learning*** problems
- We proposed PeerPL, a weakly supervised policy learning framework to unify a series of RL/BC problems with low-quality supervision signals
    - RL with perturbed reward
    - BC with imperfect demonstrations
    - Policy Co-Training (Hybrid RL + BC)
- Our method is theoretically guaranteed to recover the optimal policy with sufficient weak supervision signals.