On the Utility of Learning about Humans for Human-AI Coordination

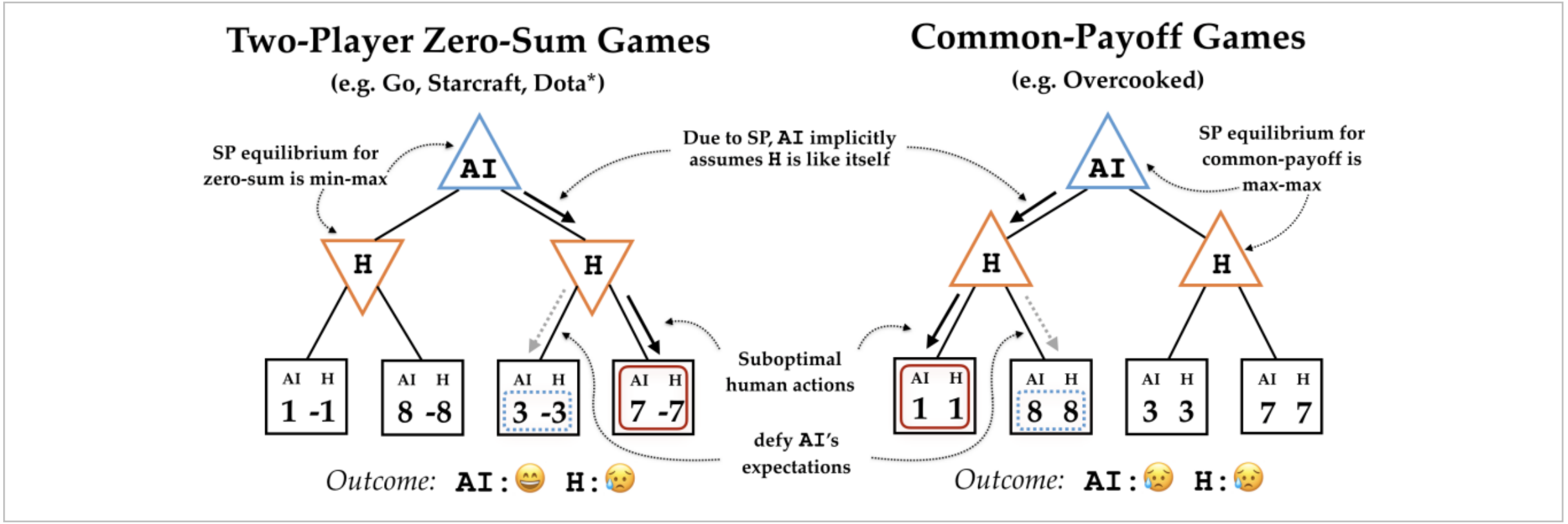
Micah Carroll[1], Rohin Shah[1], Mark K. Ho[2], Thomas L. Griffiths[2], Sanjit A. Seshia[1], Pieter Abbeel[1], Anca Dragon[1]
[1]UC Berkeley, [2]Princeton University

IIE8557-01 동적계획법과 강화학습

경영과학연구실 전재현

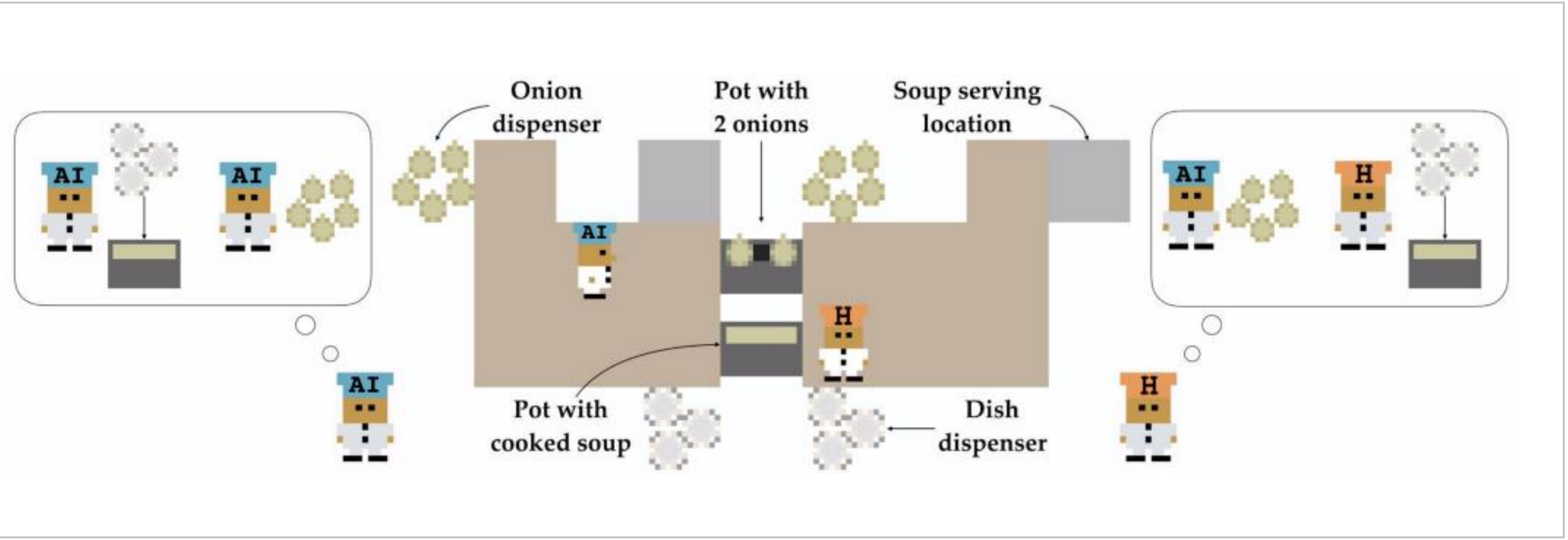
The impact of incorrect expectations

- 일반적인 게임에서 학습된 AI는 상대방이 어떤 행동을 할지를 예측하고 그에 대한 최선의 play를 함
- 하지만 협력을 해야 하는 게임에서 일반적인 게임과 같은 방식으로 학습을 한다면 오히려 최악의 결과가 나오기도 함



Overcooked game

- 두 player가 협력하여 요리를 하는 game
- 이동, 재료손질, 요리, 서빙 등 다양한 action을 할 수 있음
- 현재 전체적인 상태도 중요하지만, 다른 player가 어떤 행동을 하려고 하는지 또한 중요함

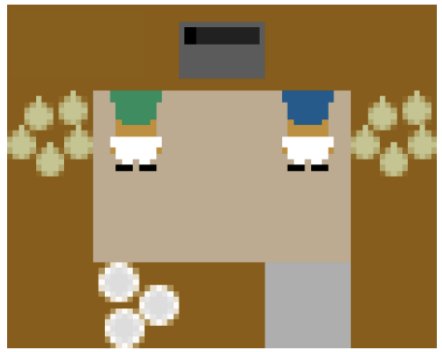


AI가 인간과 협력하는 경우
인간의 optimal하지 못한 행동에 의해
좋지 않은 결과가 나올 수 있음

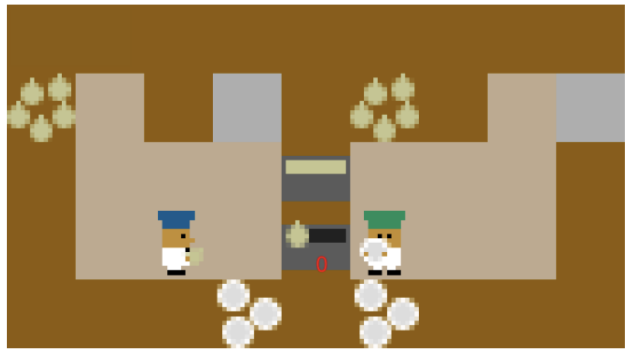
Self-play와 Behavior Cloning을 함께 활용하여 인간의 행동에 적응 가능하도록 함

Environments

- Coordination이 어려운 5개의 서로 다른 게임 상황을 만듦
- Agent들은 현재 상황에서 어떤 행동을 해야 하는지 판단할 때, 다른 agent가 무엇을 하고 있는지를 인식해야 함
- 요리가 1개 완성되어 서빙이 완료되면 일정 reward를 얻음

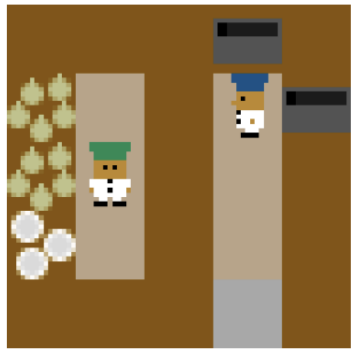


1. Cramped Room

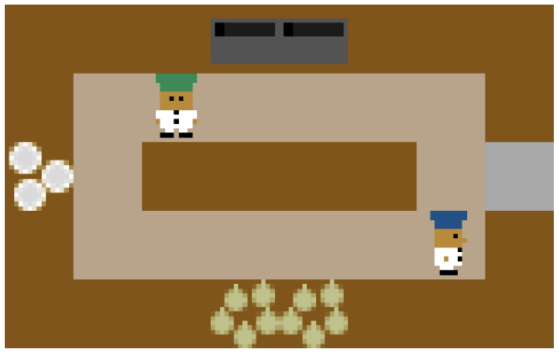


2. Asymmetric Advantages

3. Coordination Ring



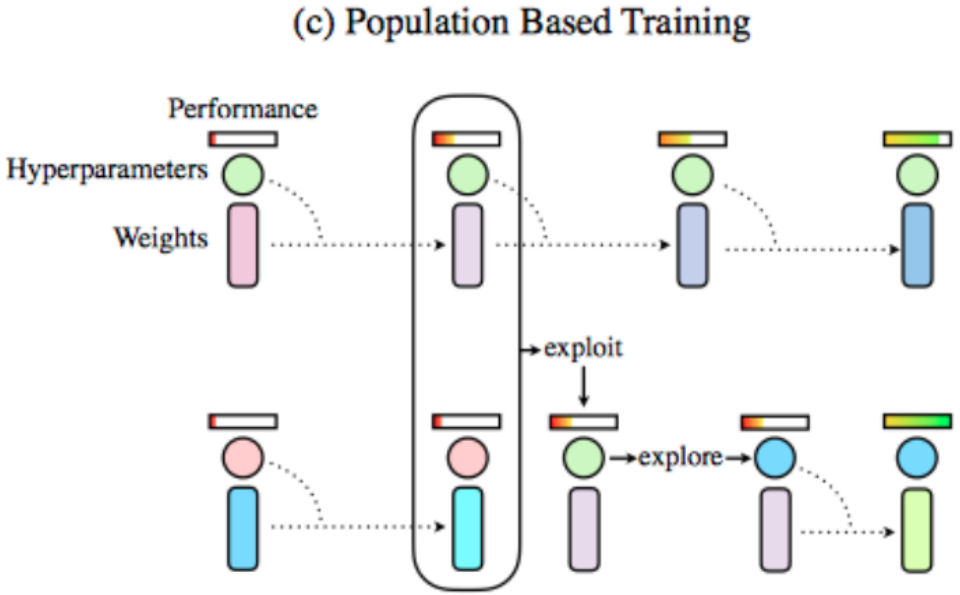
4. Forced Coordination



5. Counter Circuit

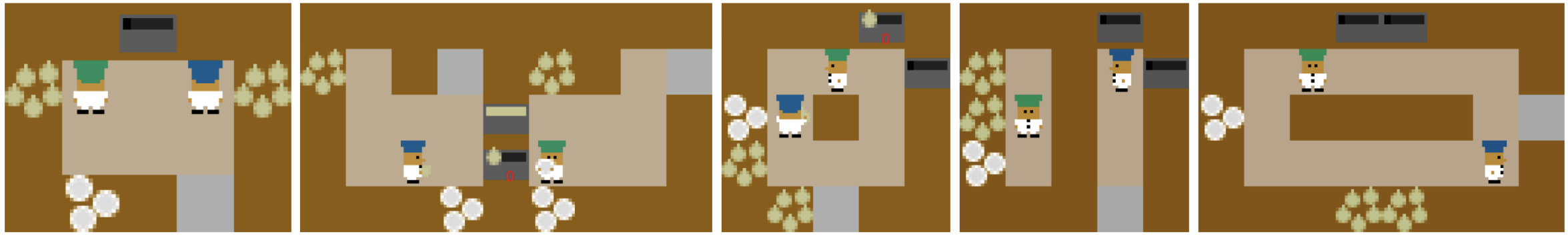
Agents

- Agents for self-play : PPO를 사용하여 에이전트를 학습시키고, 이를 self-play(SP)와 PBT 방법으로 훈련
- Agents for humans(PPO_{BC}) : PPO로 학습된 모델을 self-play로 좀 더 개선시킨 후, BC모델을 이용하여 인간의 행동에 robust하도록 함



Experiments

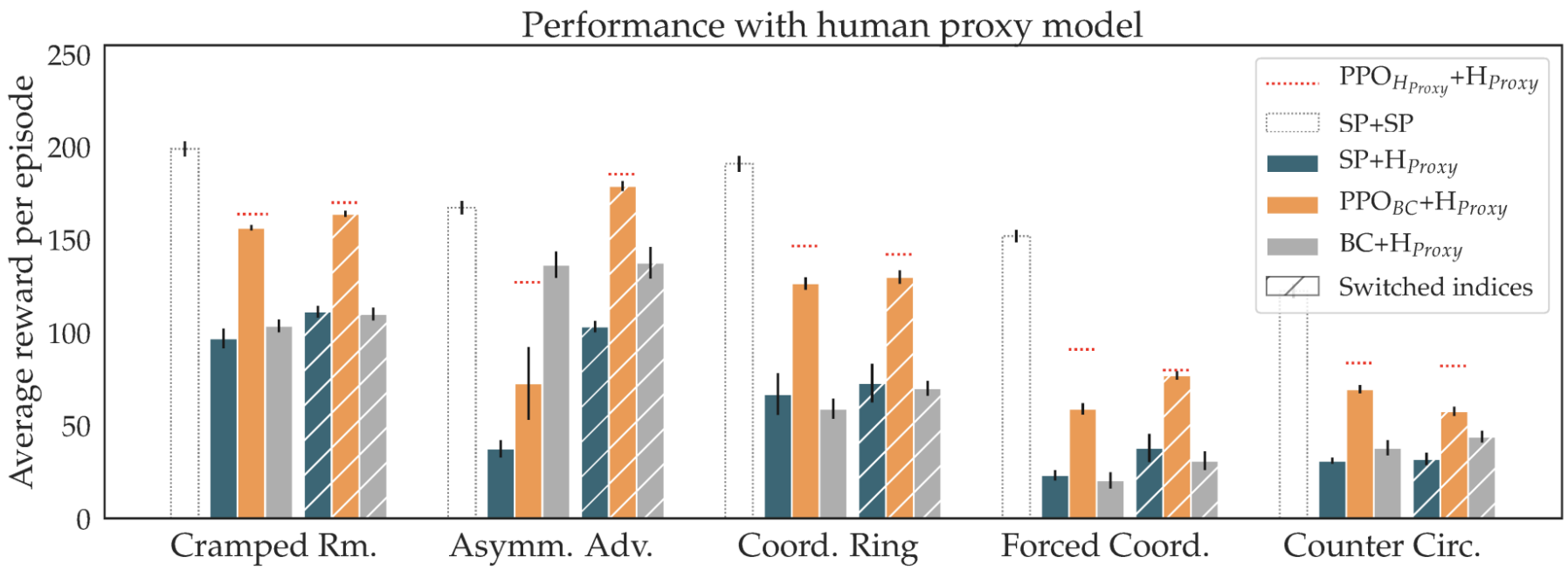
- 요리를 1개 완성하면 20의 reward를 얻음
- 400time step동안의 누적 reward를 측정함
- Agents for self-play와 Agents for humans(PPO_{BC}) 중 어떤 agent가 인간과의 협력에서 뛰어난 성능을 보여주는지를 보고자 함



1. Cramped Room
2. Asymmetric Advantages
3. Coordination Ring
4. Forced Coordination
5. Counter Circuit

Performance with human proxy model(self-play)

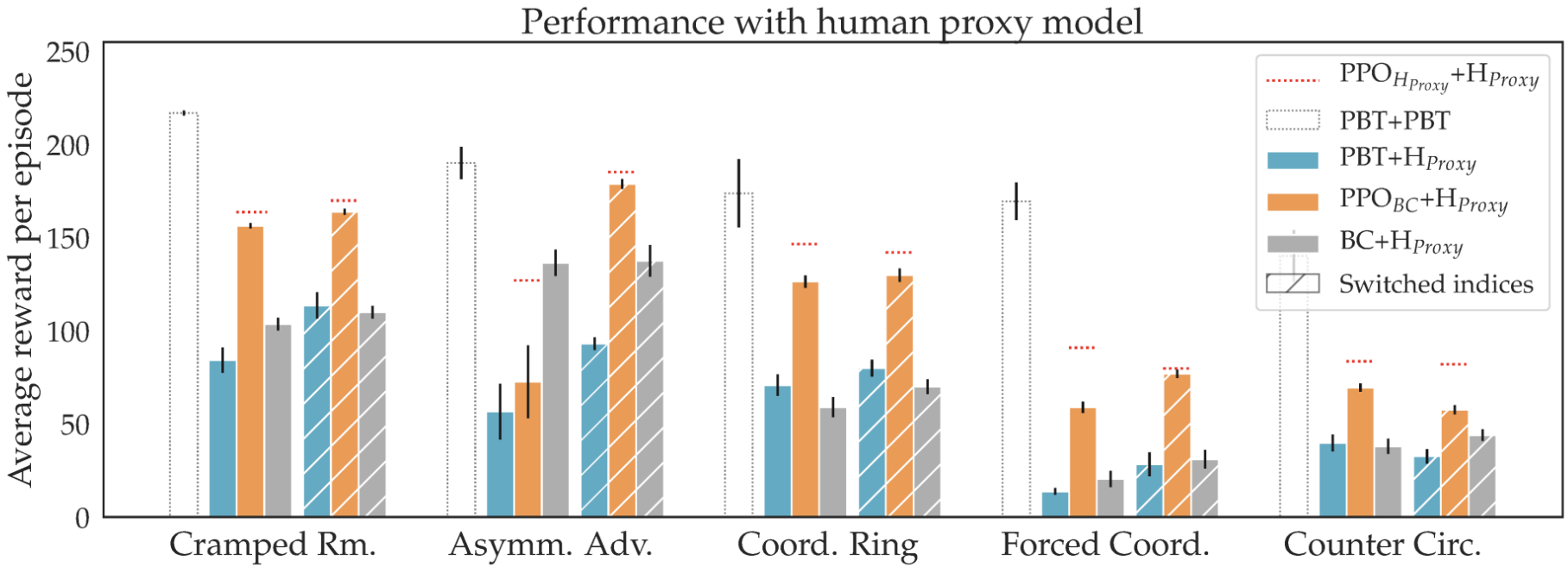
- SP+SP model의 reward가 모든 경우에 대해서 가장 큰 것을 볼 수 있음
- 하지만 SP + H_{proxy} 의 경우에 reward가 SP+SP에 비해 현저히 낮아지는 것을 볼 수 있음
- 오히려 PPO_{BC} + H_{proxy}가 대부분의 경우에 reward가 가장 높은 것을 볼 수 있음



(a) Comparison with agents trained in self-play.

Performance with human proxy model(PBT)

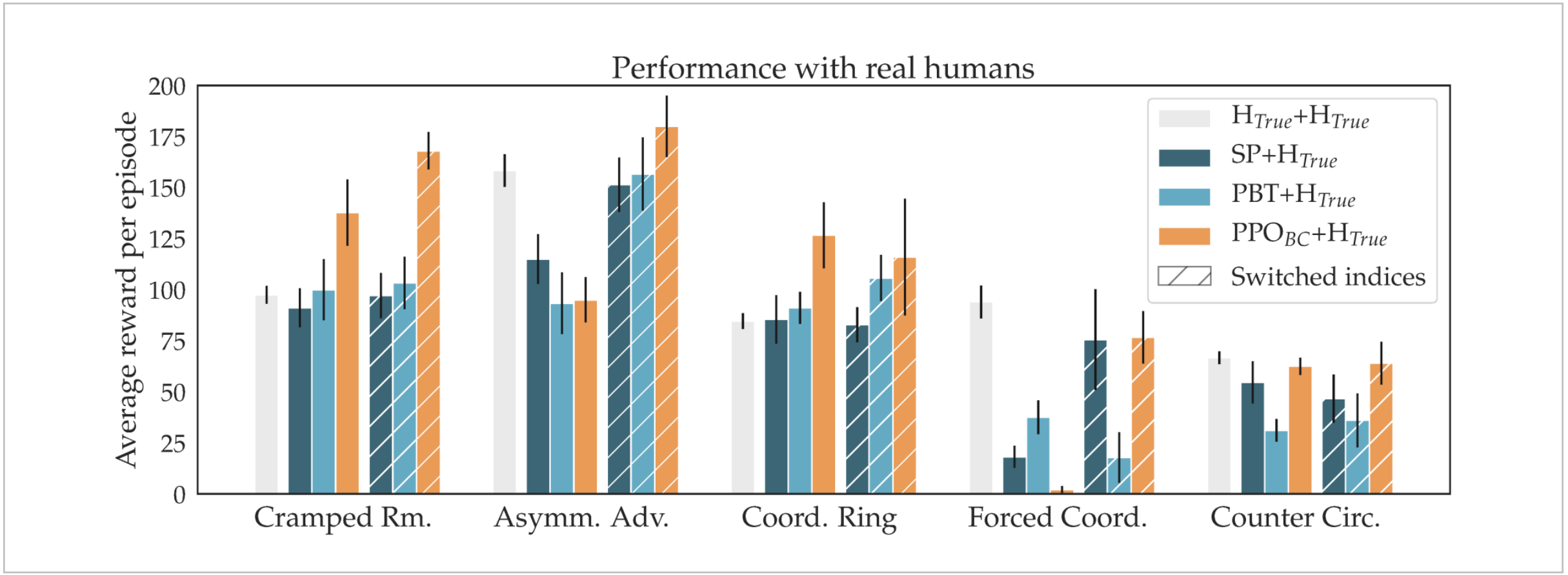
- PBT+PBT model의 reward가 모든 경우에 대해서 가장 큰 것을 볼 수 있음
- 하지만 PBT또한 SP와 마찬가지로 H_{proxy} 와 함께 play하는 경우에 reward가 PBT+PBT에 비해 현저히 낮아지는 것을 볼 수 있음
- 마찬가지로 PPO_{BC} + H_{proxy}가 대부분의 경우에 reward가 가장 높은 것을 볼 수 있음



(b) Comparison with agents trained via PBT.

Performance with real humans

- 60명의 사람을 대상으로 SP, PBT, PPO_{BC}와 함께 게임을 진행함
- 앞선 실험에서의 결과와 비슷하게 PPO_{BC} + H_{True}의 reward가 대부분 큰 것을 볼 수 있음



Conclusions

- 일반적인 DRL알고리즘을 통해 훈련된 에이전트들은 그들간의 협업에서는 뛰어난 성능을 보여주지만, 인간과 협업하는 데에 있어서는 어려움을 겪음
- 인간이 잘못된 행동을 하는 경우에 self-play를 통해 학습된 에이전트들은 해당 상황에 대해 대처를 하지 못하는 경향을 보임
- Human model과 잘 작동하도록 설계된 에이전트들이 단순한 방식으로 훨씬 더 나은 성능을 보임
- 인간과 협업하는 모델을 만들 때는 일반적인 강화학습과는 다른 방법이 필요해 보임

Q & A