# Learning to Collaborate in Markov Decision Processes

**Goran Radanovic**, Rati Devidze,
David C. Parkes, Adish Singla

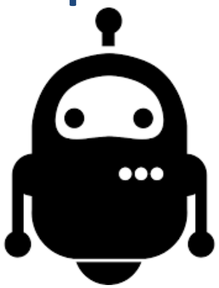
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

# Motivation: Human-AI Collaboration

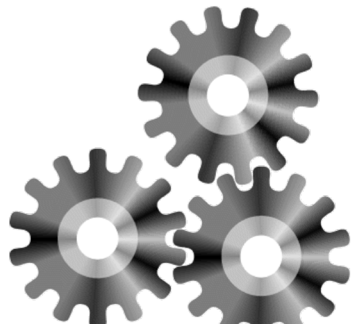
**Example setting**

**Helper-AI**



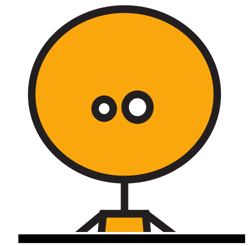
**Agent A1**

**Commits to policy $\pi_1$**



**Task**

**Human**



**Agent A2**

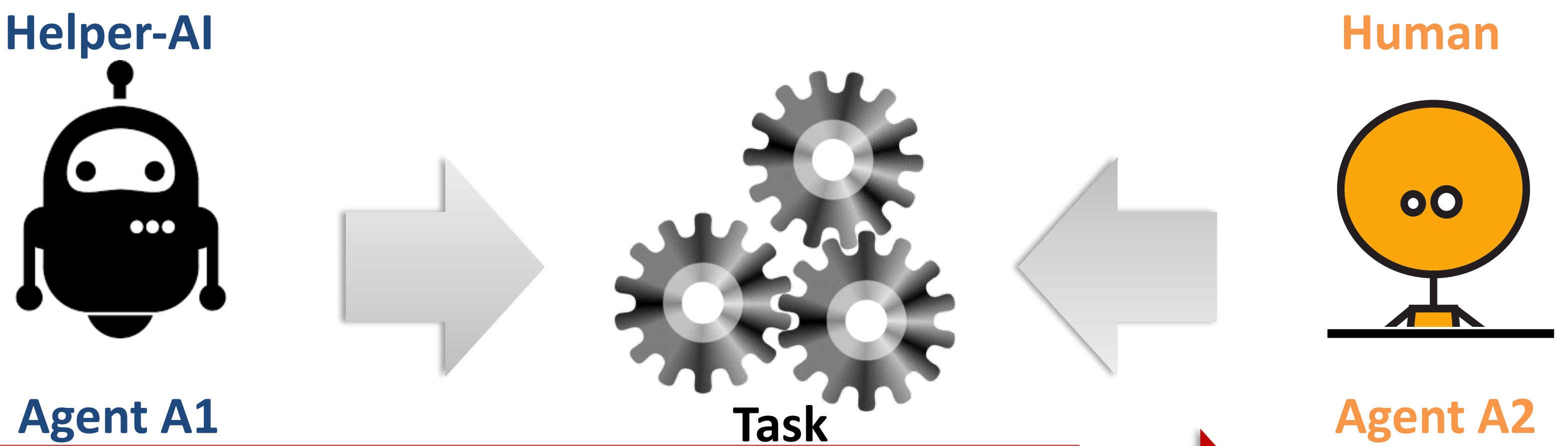
**(Best) responds to $\pi_1$**

**Behavioral differences**

Agents have different models of the world

**[Dimitrakakis et al., NIPS 2017]**

# Motivation: Human-AI Collaboration




Humans change/adapt their behavior over time.

**$\pi_2$ changes over time**

Can we utilize learning to adopt a good policy for A1 despite the changing behavior of A2, without detailing A2's learning dynamics?

# Formal Model: Two-agent MDP

- *Episodic* two-agent MDP with *commitments*

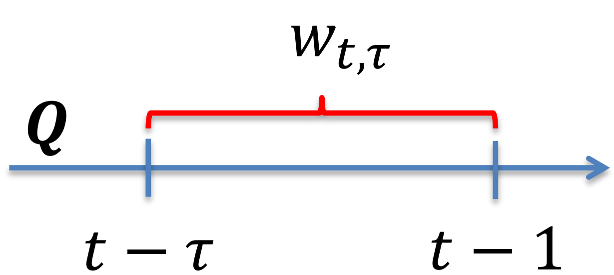
- Goal: design a learning algorithm for A1 that achieves a sublinear regret
  - Implies near optimality for *smooth* MDPs

# Experts with Double Recency Bias

- Based on experts in MDPs:
  - Assign an experts algorithm to each state
  - Use $\boldsymbol{Q}$ values as experts' losses

**[Even-Dar et al., NIPS 2005]**

- Introduce double recency bias




**Recency windowing**

$$\pi_t = \frac{1}{\Gamma}\sum_{\tau=1}^{\Gamma} w_{t,\tau}$$

**Recency modulation**

# Main Results (Informally)

**Theorem:** The regret or ExpDRBias decays as $O(T^{\max\left\{1-\frac{3\cdot\alpha}{7},\frac{1}{4}\right\}})$, provided that the *magnitude change* of A2's policy is $O(\,T^{-\alpha})$.

**Theorem:** Assume that the *magnitude change* of A2's policy is $\Omega(1)$. Then achieving a sublinear regret is at least as hard as *learning parity with noise*.

# Thank you!

- Visit me at the poster session!