

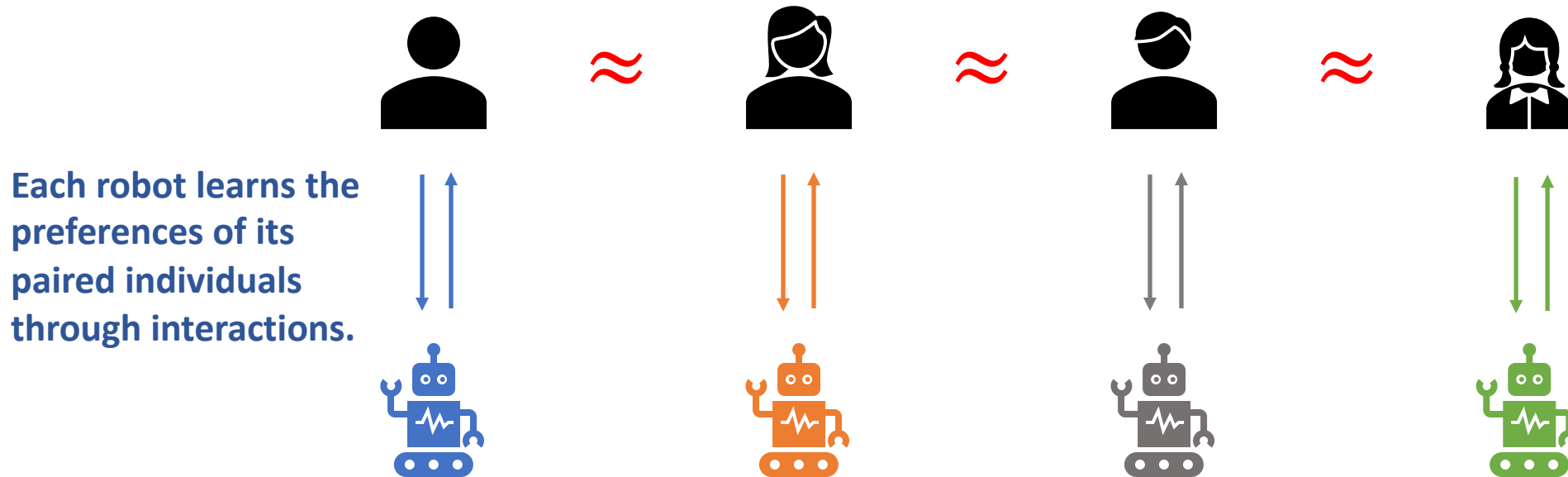
Stochastic Multi-Player Bandit Learning from Player-Dependent Feedback

Zhi Wang*, Manish Kumar Singh, Chicheng Zhang,
Laurel D. Riek, Kamalika Chaudhuri

UC San Diego
JACOBS SCHOOL OF ENGINEERING

 **THE UNIVERSITY
OF ARIZONA**

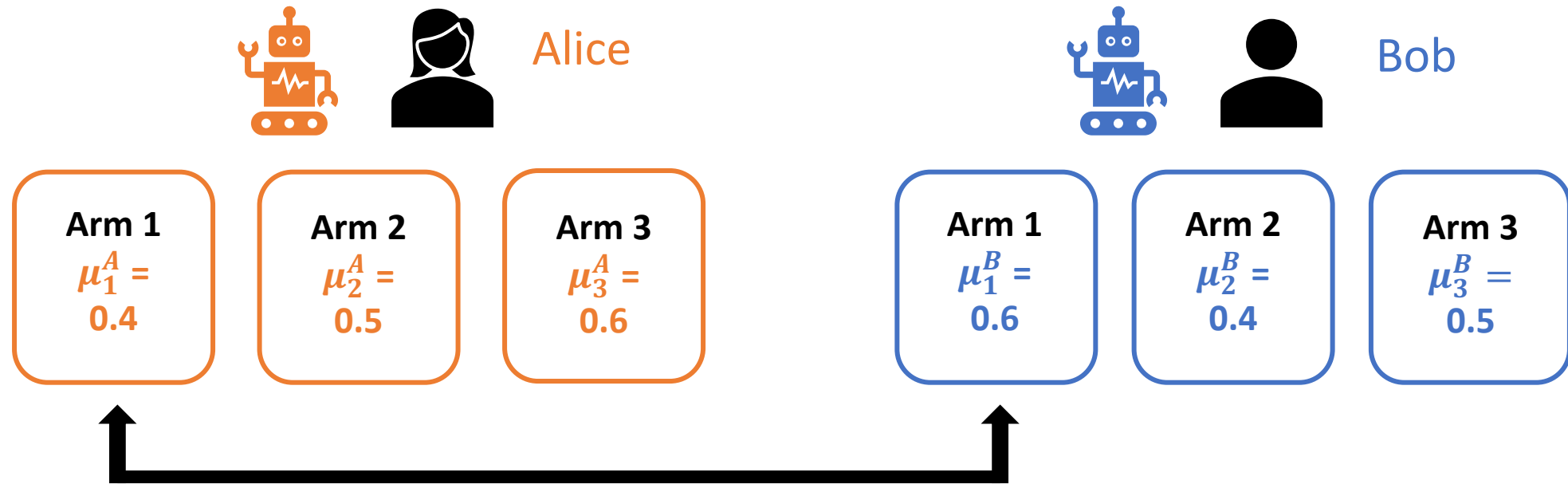
Heterogenous Multi-Task Online Learning



- A group of assistive robots deployed to provide personalized healthcare services.
- Question: If the robots receive **similar yet nonidentical** feedback, how do they learn to perform their respective tasks faster in an **online** learning setting?

(Stochastic) ε -Multi-Player Multi-Armed Bandit

- A set of M players (robots) concurrently interact with K arms.



$\forall p, q \in [M], \max_{i \in [K]} |\mu_i^p - \mu_i^q| \leq \varepsilon \longrightarrow \varepsilon \in [0,1]$ discrepancy parameter

- In each round, every player pulls one arm and the players share information at the end of each round.

Results and Future Work

Approach: *Adaptively and robustly* aggregate rewards shared by other players to construct “high probability” confidence intervals

Results:

- When ε is sufficiently small, we can obtain a problem-dependent upper bound on expected collective regret (sum of each player’s regret) that has an *inconsiderable* dependence on M ;
- **Fall-back guarantee**: for large ε ’s, our performance guarantee is never (by a constant factor) worse than that of running UCB-1 for each player individually.

Future Directions: unknown ε , extension to linear contextual bandits, etc.