



Achieving Alignment Through Adaptive Play: Helping Optimize Objectives Without Observing Them

STUDENTS: Jason T. Isa

Problem: Multi-Agent Alignment Without Seeing Objective

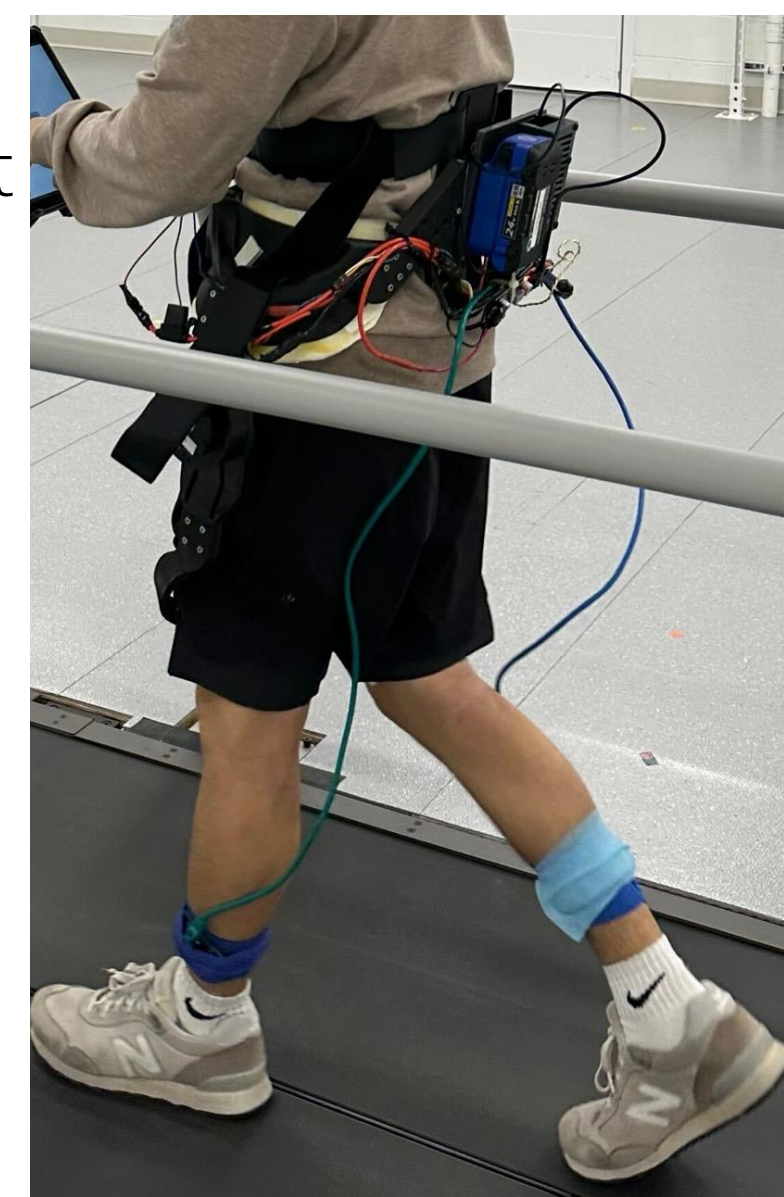
In many human-machine and multi-agent AI systems, one agent optimizes a **private objective**, while another agent seeks to assist **without observing it**.

We study repeated interactions where:

- Agent 1 (optimizing agent) minimizes a hidden cost $f(x, u)$
- Agent 2 (helper) observes only actions
- Agent 2 has no access to cost values, gradients, or preferences

Question:

Can a helper provably align with a hidden objective using only observed adaptive behavior?



Example: Human-Exoskeleton Assistance

Human adapts their walking to minimize **metabolic effort**

Exoskeleton observes the human's movements but **cannot measure the human's cost directly**, and must learn how to assist through interactions.

Algorithm

Algorithm 1 Two-timescale affine-response method (Uniform Sampling)

- 1: Initialize $(x_0, u_0) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.
- 2: **for** $k = 1, \dots, K$ **do**
- 3: Sample L_k uniformly from \mathcal{L} and set $L_k = L_{i_k}$.
- 4: Define $\pi_k(x) = u_{k-1} + \nu_k L_k (x - x_{k-1})$.
- 5: Set $x_{k,0} \leftarrow x_{k-1}$ and $u_{k,0} \leftarrow u_{k-1}$.
- 6: **for** $t = 0, \dots, T - 1$ **do**
- 7: $x_{k,t+1} \leftarrow \mathcal{A}(x_{k,t}; L_k, u_{k-1}, x_{k-1})$
- 8: $u_{k,t+1} \leftarrow \pi_k(x_{k,t+1}) \triangleright$ policy-implied update
- 9: **end for**
- 10: $(x_k, u_k) \leftarrow (x_{k,T}, u_{k,T})$.
- 11: **end for**

Outer loop (Helper / Agent 2)

At each epoch k , the helper **selects a subspace** by sampling $L_k \in \{L_1, \dots, L_N\}$ and commits to a **fixed affine policy**

$$\pi_k(x) = u_{k-1} + \nu_k L_k (x - x_{k-1})$$

This constrains the the joint action (x, u) to an **affine subspace** through the previous iterate, allowing the helper to steer learning **without observing costs**.

Inner loop (Optimizer / Agent 1)

With π_k fixed, Agent 1 interacts for T steps to approximately minimize the induced objective

$$\phi_k(x) = f(x, \pi_k(x))$$

using its own update rule \mathcal{A} (e.g., best response, SGD, PPO). The resulting iterate (x_k, u_k) remains on the helper-imposed subspace and becomes the starting point for the next epoch.

Algorithm Design Parameters

Parameter	Controls	Effect
Response Gain ν_k	Scales the helper's affine response	Small \rightarrow slow progress Large \rightarrow amplifies noise and increases noise floor
Epoch Length T	Number of inner-loop adaptation steps	Longer epochs improve inner optimization and accelerates outer convergence, but can result in more noise per epoch

Convergence

Assumptions

- (1) f is smooth
- (2) PL Condition
- (3) Subspace alignment

Main Result

Without observing the costs or gradients, the helper achieves linear convergence by repeatedly inducing aligned subspace responses from the optimizing agent.

Stochastic Optimizing Agent

$$\mathbb{E}[f(z_k) - f^*] \leq \rho^k (f(z_0) - f^*) + \varepsilon_\infty$$

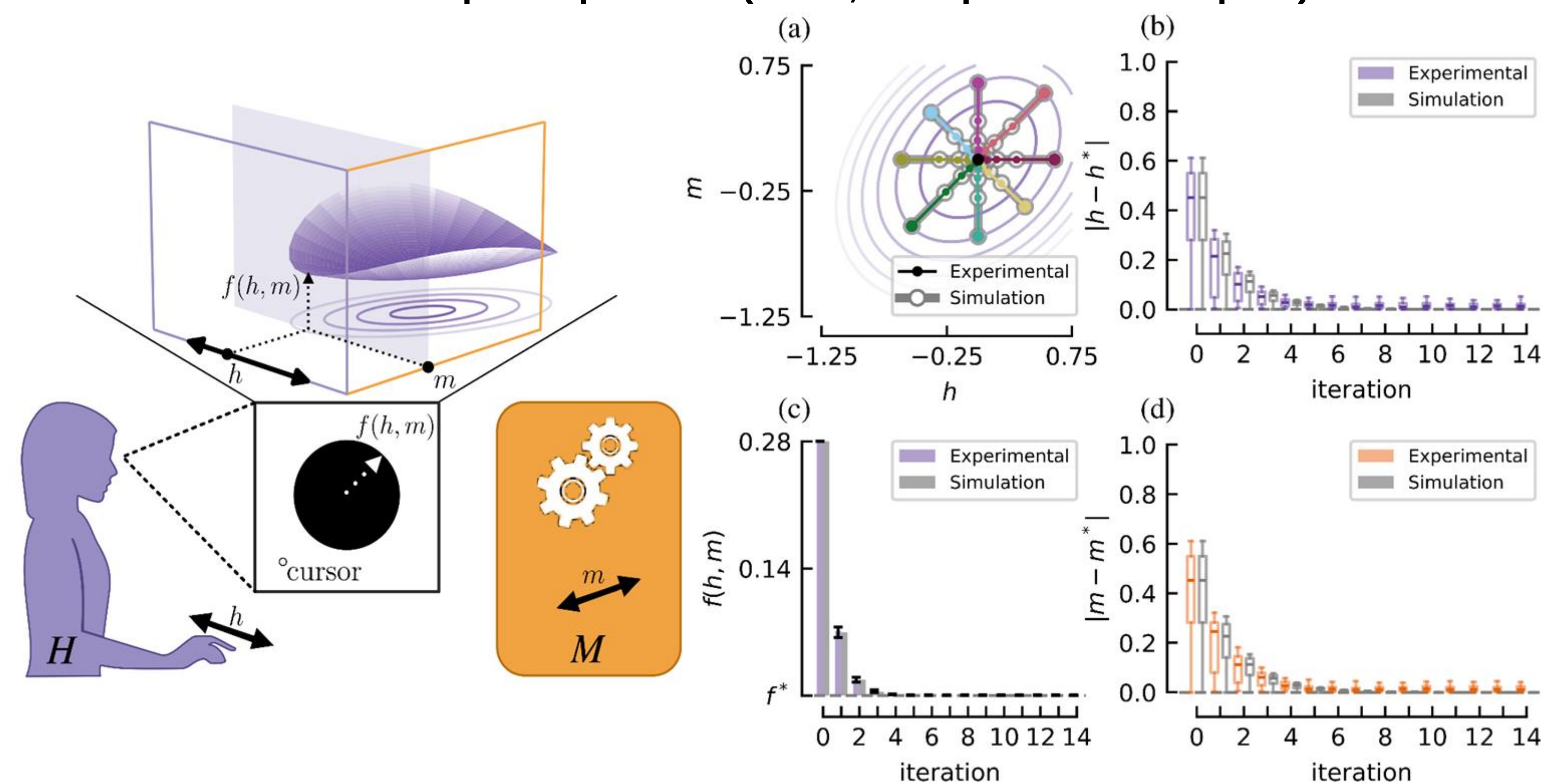
Deterministic Optimizing Agent

$$\mathbb{E}[f(z_k) - f^*] \leq \rho^k (f(z_0) - f^*)$$

where $\rho \in (0,1)$ and ε_∞ is the steady-state noise floor

Empirical Results

Human Participant Experiments (n = 80; n = 10 per initialization point)



Two-Agent Balancing Cart Pole Experiment

