Effect of Adaptation Rate and Cost Display in a Human-AI Interaction Game

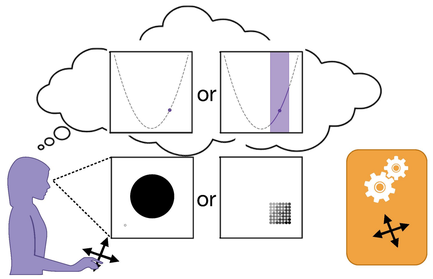
STUDENTS: Jason T. Isa, Bohan Wu, Qirui Wang, Yilin Zhang

Introduction

- As interactions between humans and AI become more prevalent, it is critical to have better predictors of human behavior in these interactions to create more safe and socially beneficial outcomes.
- Previous work has shown that AI can unilaterally change their adaptive algorithm to drive the outcome of a simple Human-AI game to distinctly different game theoretic equilibria [1].

Problem

- Without changing the AI agent's adaptive algorithm, can we influence the Human-AI game outcome by changing the Human's feedback information and model these behavior changes within a game theoretic model?



Game Theoretic Equilibria

Nash Equilibrium

$$h^{\mathrm{N}} = \arg\min_{h \in \mathcal{H}} c_H(h, m^{\mathrm{N}}),$$
$$m^{\mathrm{N}} = \arg\min_{m \in \mathcal{M}} c_M(h^{\mathrm{N}}, m)$$

Stackelberg Equilibrium

$$h^{\mathrm{S}} = \arg\min_{h \in \mathcal{H}} \{c_H(h, m^{\mathrm{S}}) \mid m^{\mathrm{S}} = \arg\min_{m} c_M(h, m)\},$$
$$m^{\mathrm{S}} = \arg\min_{m \in \mathcal{M}} \{c_M(h^{\mathrm{S}}, m)\}.$$

Experimental Algorithm

Algorithm 1 Experiments with human subjects

Require: initial $(h_{(0,:)}, m_{(0,:)})$, parameter $\gamma > 0$
  for $t = 0, 1, \ldots, T-1$ **do**
    switch Experiment **do**
      case 1
        display $c_H(h_{(t,:)}, m_{(t,:)})$
      case 2
        display $\{c_H(h, m_{(t,:)}) \mid h \in B_\epsilon(h_{(t,:)})\}$

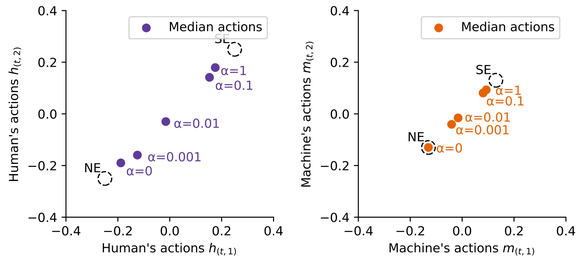
$h_{(t+1,:)} =$ `manual_input(t)` ▷ Human input
$m_{(t+1,:)} = m_t - \alpha \frac{\partial c_M}{\partial m}(h_{(t,:)}, m_{(t,:)})$ ▷ AI update
  end for

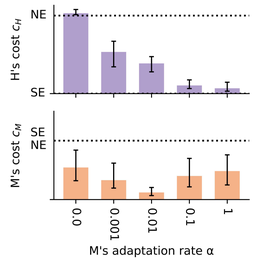
Where both Human and AI cost functions were defined as quadratic functions,

$$c_H(h, m) = \frac{1}{2} h^\top A_H h + h^\top B_H m + \frac{1}{2} m^\top D_H m + h^\top a_H + m^\top b_H$$
$$c_M(h, m) = \frac{1}{2} m^\top A_M m + m^\top B_M h + \frac{1}{2} h^\top D_M h + m^\top a_M + h^\top b_M$$

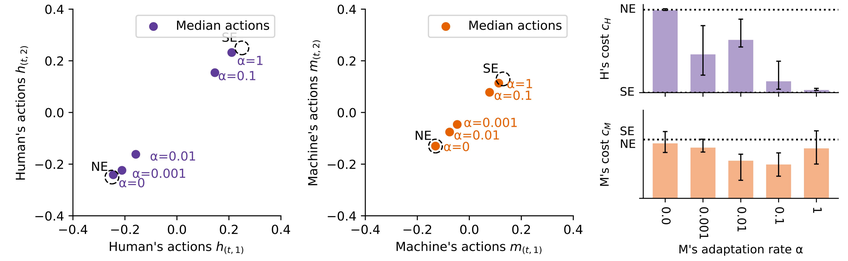
Empirical Results

Cost Display

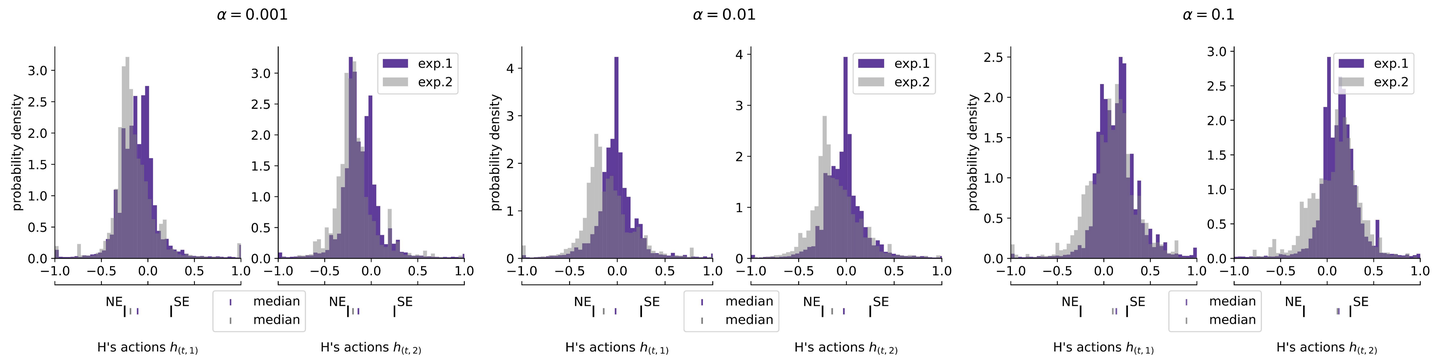




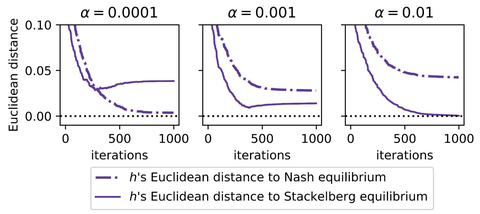
Cost Landscape Display

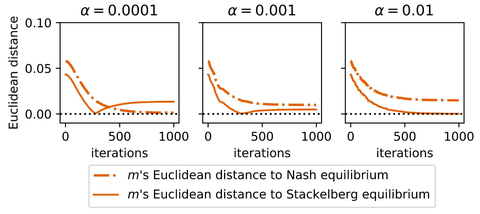


Cost Display vs Cost Landscape Display



Simulations





- Simulations of learning dynamics of a game with action vector dimensions of 64 inputs for Agent H and 128 inputs for Agent M.
- Agent H adapted their actions according to a multi-point gradient estimation.
- Agent M adapted their actions according to the same gradient descent rules as the AI in the human experiments.

Simulation Results

- For slow adaptation rates, the actions shift towards the Nash equilibrium, while for fast adaptation rates, the actions shift towards the Stackelberg equilibrium, matching the experiments.

Conclusion

- Our work gives a simple predictive model for certain tasks, which we have empirically validated, and ultimately for influencing and shaping behavior in Human-AI interaction tasks.
- Our work looks at adaptation in terms of a gradient or gradient like quantities, so this could possibly be integrated into policy gradient and active gradient methods in reinforcement learning.

References

1. Benjamin J. Chasnov, Lillian J. Ratliff, and Samuel A. Burden. 2023. Human adaptation to adaptive machines converges to game-theoretic equilibria. (2023). arXiv:2305.01124 [cs.AI]

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISERS: Samuel A. Burden, Lillian J. Ratliff
COLLABORATOR: Benjamin J. Chasnov

Questions?
email: jisa@uw.edu