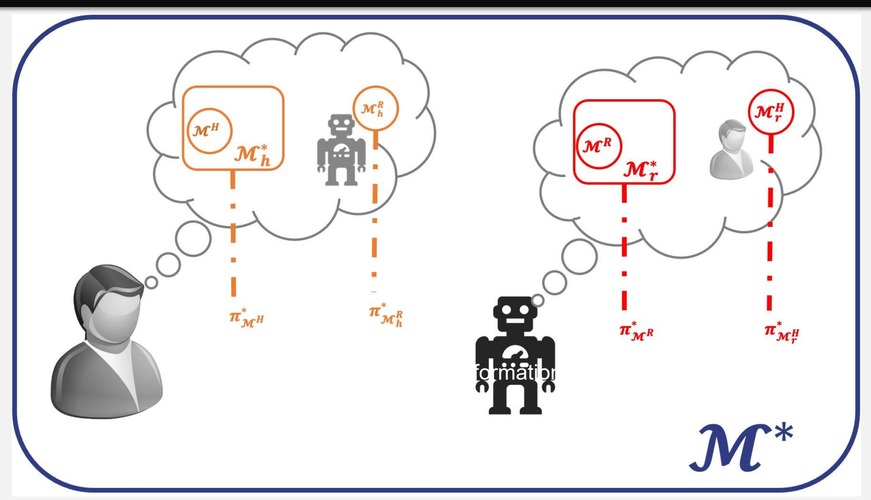
A Mental-Model Centric Landscape of Human-AI Symbiosis

Zahra Zahedi, Sarath Sreedharan, Subbarao Kambhampati
School Computing and AI, Arizona State University

Introduction

- There is a lack of a unifying framework that can make sense of the diverse challenges and tools used in the field of Human AI interaction.
- This paper presents a new framework called GHAI (Generalized Human-Aware interaction) that addresses this gap.
- GHAI not only allows for scenarios where the human may be an active participant but also introduces the notion of a true task model that captures the true joint task specification of both the human and AI agent.
- Our model separates the agent's belief about the true underlying joint task, from the other agent's perception of it.
- As discussed in the paper, this framework is successful in unifying works from several communities to address different modes of human-AI interaction including human-aware planning, human-AI cohabitation, human-robot interaction, human-in-the-loop ML, and human-AI symbiosis.

Generalized Human-Aware Interaction



Agents– Our central modeling captures the interaction between a human (H) and AI agent (R)

Model– A task model ($M_.^.$) is any mathematical model that encodes among other things, an entities beliefs about task objectives, state of the world and how the world may evolve on its own or in response to an agent action. (e.g. MDPs, differential system equations or symbolic models like PDDL)

Task models $M_.^*$– This model captures the entirety of the task. It consists of all of the actions, objectives, preferences of both agents and additional facts about the world state that may not fall into the purview of the individual agent models. We have the ground truth model $M^*(M_*^*)$, and the task models maintained by each of the agents (M_h^* and M_r^*).

Model of each agent M^R and M^H – This is the model each agent ascribes to themselves. This determines what actions each agent believes they could perform and the objectives and preferences they are trying to satisfy. These models are also part of each agent's beliefs about what the true task model is, i.e., M^R is part of M_r^* and M^H is part of M_h^*.

One agent's model of the other M_h^R and M_r^H – These consist of what one agent believes the model of the other agent is, i.e, M_h^R is the human's belief about M^R, and M_r^H is the AI agent's belief about M^H.

Decisions – We will generally denote decisions as $\pi_.$ with the subscript corresponding to the model from which it is derived.
This is not limited to a specific decision type and depending on the exact problem the space of decisions could vary from single shot labels one may associate with a classification task to a policy to be carried out by an embodied agent.

Landscape of Human-AI Symbiosis

- **Model Information from AI Agent's End (M_r^*) --** works that mostly use the AI models to update models held by the human.
  We can further categorize works based on
  - the human models that will be updated (i.e., M_{h}^R or M^H)
  - Whether or not the AI agent makes use of a mental model of the human (M_r^H) to generate the required information

References	Model information from $\mathcal{M}_r^*$		
	$\mathcal{M}_r^H$ Used	Updating $\mathcal{M}^H$	Updating $\mathcal{M}_h^R$
[6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 24; 25; 26; 27; 28; 29; 31; 19; 20; 21; 22; 23; 30]	✗	✗	✓
[38; 2; 18]	✓	✗	✓
[37]	✗	✓	✓
[32; 33; 34; 35; 36]	✓	✓	✗
[39]	✓	✓	✓

- **Model Information from Human's End (M_h^*) --** works that use the Human mental models to update AI agent models.
  We have model information coming from M_h^* that is used to
  - update M^R, M_r^H , or both
  - The human might use M_h^R to select this information

References	Model information from $\mathcal{M}_h^*$		
	$\mathcal{M}_h^R$ Used	Updating $\mathcal{M}^R$	Updating $\mathcal{M}_r^H$
[54; 34; 33; 36; 35; 55; 56]	✗	✗	✓
[40; 41; 42; 43; 41; 45; 46; 47; 51; 52]	✗	✓	✗
[48; 49; 13; 57]	✓	✓	✗
[50; 53; 58]	✓	✓	✓

- **Model Following Behavior --** how agents may choose decisions that take into account multiple models.
  In particular, we have
  - Multi model alignment
    - M_r^* and M_h^R
    - M_r^* and M_r^H
    - M_r^* and M_h^*
    - M_h^* and M_h^R
  - Response seeking behavior

Model Following Behavior				
Multi-Model Alignment				Response-Seeking Behavior
$\mathcal{M}_r^*$ & $\mathcal{M}_h^R$	$\mathcal{M}_R^*$ & $\mathcal{M}_r^H$	$\mathcal{M}_r^*$ & $\mathcal{M}_h^*$	$\mathcal{M}_h^*$ & $\mathcal{M}_h^R$	$\mathcal{M}_r^H$
[59; 60; 61; 62; 63; 24; 25]	[64]	[69; 65; 66]	[67]	[68; 31; 54]

Acknowledgements

This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1- 2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19- 2-0006, and a JP Morgan AI Faculty Research grant.

Ira A. Fulton Schools of Engineering
Arizona State University

Yochan Lab
https://yochan-lab.github.io/home/