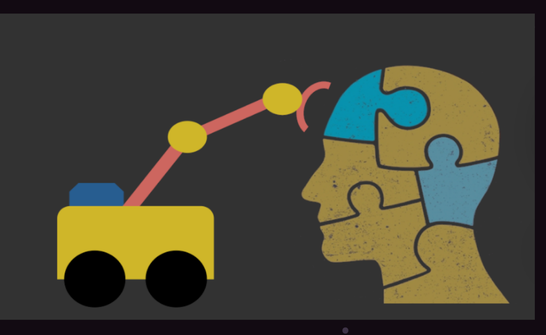
EAGER: Reconciling Model Discrepancies in Human-Robot Teams

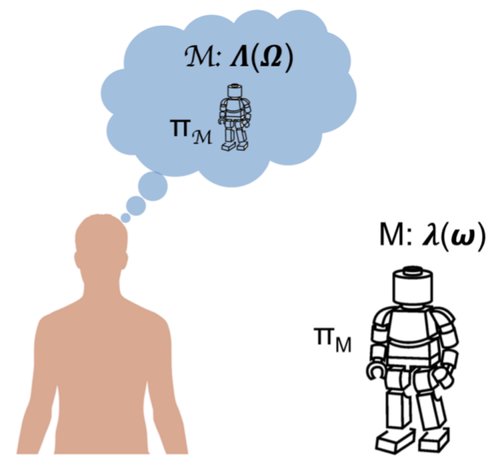


Yu ("Tony") Zhang
Arizona State University, USA
yzhan442@asu.edu

1. Motivation

- Teammates have many conscious and subconscious ***expectations*** of others in terms of their plans or behaviors
- The expected model (EM) and actual model (DM) may differ, leading to unmatched expectations, loss of situation awareness and trust

➢ This calls for general methods for ***model reconciliation***



2. Research Thrusts and Intellectual Merit

There are at least four aspects that may be reconciled (some of which have been studied previously):

- ***Goal***: reconciling between the goals of agents
- ***Behavior***: directly reconciling the plans/behaviors
- ***Domain***: reconciling domain dynamics, including initial state [*knowledge*]
- ***Cognitive model***: reconciling cognitive capabilities [*computation or inference process*]

Type	Align M with $\mathcal{M}$		Align $\mathcal{M}$ with M	
	Exp. AI		Exp. AI	Exp. AI
	Explicit	*Implicit*	*Explicit*	*Implicit*
Goal	`Rwd Lrn` [31, 33] `Pln Rec` [22, 73] `Ins Und` [17, 123]	`IRL` [5, 138]	`Inv Grd` [122] `Int Prj` [7, 56] `Gol Aug` [23]	`Lgb Pln` [34] `Pln CoE` [77] `Com Act` [75]
Behavior	`Imi Lrn` [107, 111] `HuW Pln` [21, 26]	`NA`	`Crs Trn` [94]	`NA`
Domain	`Opn Wrd` [120] `Dom Lrn` [135]	**RT 1, 2**	**RT 1, 2** `Exu Gen*` [53, 60]	`Exp Inc` [79] `Com Act` [75]
Cognition	`Met Rsn` [110] `Com Rat` [47]	`DvP` [92]	`NA`	**RT 3**

3. Technical contributions and innovation

Learning methods for domain model reconciliation

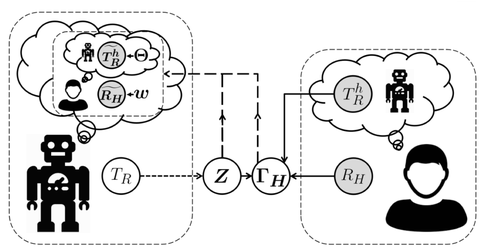
- Applying IRL may lead to learning the opposite human preferences when humans are biased!
- Generalized reward learning under biased domain dynamics (AAAI 2020)

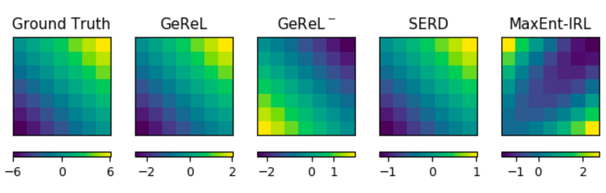
Given:

- Robot's demonstrations $\boldsymbol{Z}$;
- Human's ratings $\boldsymbol{\Gamma_H}$ for each instance in $\boldsymbol{Z}$.

To determine:

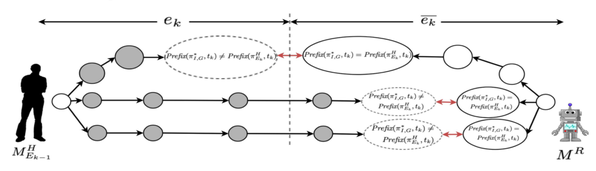
- Human's true reward function R_H;
- Human's belief T_R^h about robot's domain dynamics.

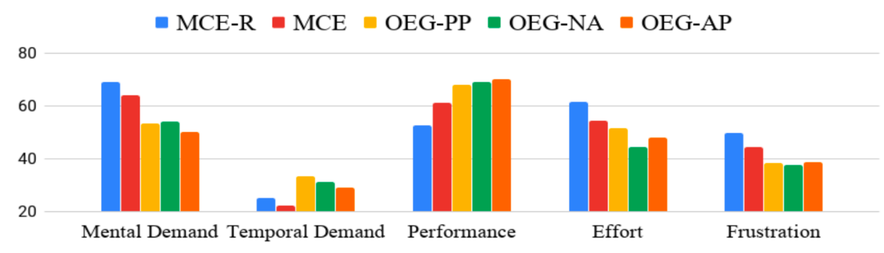




Reconciling cognitive models

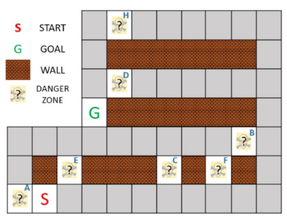
- Cognitive models are difficult to reconcile directly
- But we can reconcile them indirectly!
  - OEG (ready for IROS)!
  - Break an explanation into multiple pieces and communicate them online

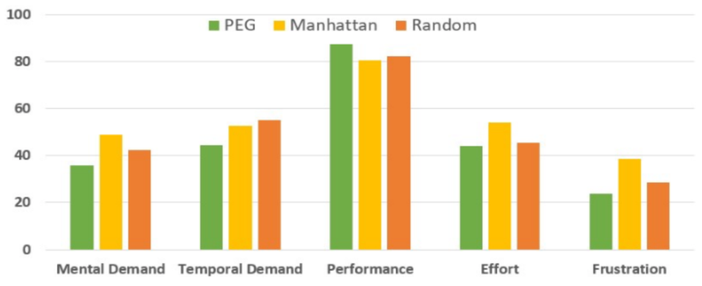




Reconciling cognitive models (cont.)

- But we can reconcile them indirectly!
  - PEG (under submission to IJCAI)
  - Order of communicating information in an explanation matters!
  - Learn the order via IRL methods





4. Broader Impact: Education

- *Keynote speaker at Intel, Chandler on "Challenges in Cognitive Human-robot Teaming"*
- *Judging for Intel ISEF, involving high school students from around the world*
- *Engineering projects for graduate student at ASU*
- *CSE 591 "Human-Aware Robotics", covering research methods developed in this research*

5. Broader Impact: Societal

- *Ubiquitous collaborative robots require robotic technologies that support human-robot teaming*
- *Safety and trust issues*
- *Co-bot technology for improving our everyday life*
- *Interpretable and explainable AI (AI explains not only its decision but also its behavior)*
- *ONR, ARO, NIH programs*

2020 National Robotics Initiative (NRI) Principal Investigators' Meeting
FEBRUARY 27 - 28, 2020 | ARLINGTON, VIRGINIA

Award ID #: IIS-1844524