Mutual Understanding in Human-Machine Teaming

Rohan Paleja
Institute of Robotics and Intelligent Machines
Georgia Institute of Technology
Atlanta, GA 30332
rohan.paleja@gatech.edu

Abstract

Collaborative robots (i.e., "cobots") and machine learning-based virtual agents are increasingly entering the human workspace with the aim of increasing productivity, enhancing safety, and improving the quality of our lives. These agents will dynamically interact with a wide variety of people in dynamic and novel contexts, increasing the prevalence of human-machine teams in healthcare, manufacturing, and search-and-rescue. In this research, we enhance the mutual understanding within a human-machine team by enabling cobots to understand heterogeneous teammates via person-specific embeddings, identifying contexts in which xAI methods can help improve team mental model alignment, and enabling cobots to effectively communicate information that supports high-performance human-machine teaming.

Introduction

The field of human-machine teaming (HMT) is concerned with understanding, design, and evaluation of machines for use by or with humans (Chen and Barnes 2014). While researchers within the fields of HMT, Human-Robot Interaction (HRI), and Explainable AI (xAI) have made many improvements to allow for higher quality human-machine interaction, three primary challenges limit our ability to effectively integrate cobots into human work environments.

First, there is significant heterogeneity across humans in their decision-making strategies, making it difficult for cobots to anticipate and adapt to their human partners. Second, humans are typically unable to understand and anticipate cobot behavior. Legible cobot behavior is necessary so the human can maintain a high level of predictability (i.e., develop a shared mental model) in HMT. Lastly, humans have a limited cognitive bandwidth and can only process a finite amount of information. Thus, cobots must be able to communicate efficiently with humans, only communicating when necessary with essential information. The aim of my thesis is thus: (1) enable cobots to understand heterogeneous end-users using white-box learning (xAI) methods, (2) identify the contexts in which such xAI methods can help improve HMT, and (3) enable cobots to efficiently communicate information that supports HMT. By tackling these key challenges, we allow for increased mutual understanding within human-machine teams.

Understanding Heterogeneous Decision-Making

Human decision-makers utilize rules-of-thumb and strategies honed over decades of apprenticeship (i.e., unique heuristics depending on experts' varied experiences and personal preferences) in solving complex sequential decision-making problems. As cobots will interact with a wide variety of people, cobots must be able to tailor their behavior to each human, a feat demonstrated in Paleja et al. (2020). Early work (Sammut et al. 2002) found that pilots executing the same flight plan created such variance in the data as to make it more practical to learn from a single pilot and disregard the remaining data.

In Paleja et al. (2020); Paleja and Gombolay (2019), we scale beyond the power of a single decision-maker and are able to learn from multiple, heterogeneous decision-makers by developing a novel, data-efficient apprenticeship learning framework. Here, I design an architecture that serves as a function approximator specifically designed for sparsity to afford easy "discretization" into a Boolean decision tree after training as well as the ability to leverage variational inference to tease out each demonstrator's unique decision-making criteria. Our framework utilizes person-specific embeddings, learned through backpropagation, which enables the apprenticeship learner to automatically adapt to a person's unique characteristics while simultaneously leveraging any homogeneity that exists within the data.

In this research, I compare our novel learning from heterogeneous demonstration framework to several baselines and find that we not only outperform state-of-the-art apprenticeship learning frameworks ($+51\%$ and $+11\%$ imitation accuracy across scheduling problems in synthetic and real-world domains, respectively), but we are also able to do so with a framework that can be discretized into a human-readable form. Utilizing our framework (Paleja and Gombolay 2019; Paleja et al. 2020), cobots can gain an implicit understanding of their human teammate's behavior via an inferred representation of the teammate's policy and insight into the teammate's unique characteristics via a person-specific embedding, allowing cobots to collaborate with a wide variety of

humans while tailoring to the needs of their unique team.

Utility of xAI in HMT

The next step in developing high-performance HMT is to gain insight into the interaction between teammates in a HMT in the presence of interpretable cobot policies. Interpretable policies provide a human teammate insight into the AI's rationale, strengths and weaknesses, and expected behavior. Without this insight, it is impossible for large-scale cobot adoption in safety-critical and legally-regulated domains (Doshi-Velez and Kim 2017). In Paleja et al. (2020), I develop a Personalized Neural Tree (PNT) model that learns a model of heterogeneous user decision-making using counterfactual reasoning via pairwise comparisons. I conducted a novel user study to assess the interpretability of our framework in comparison to neural networks, designing an online questionnaire that asks users to make predictions (i.e., compute the model output given inputs) following the respective models. We find that our discrete trees are more interpretable, easier to simulate, and quicker to validate than neural networks. This finding provides support that users will be able to utilize our high-performance tree-based model in decision-tracking of cobot behavior.

In Paleja et al. (2021), I look at the utility of interpretable policy abstractions in real-time ad hoc human-machine teaming. I design a complex HMT scenario within the Minecraft domain, where a human and AI must work together to build a multi-level house. During the interaction, the cobot policy may be displayed to the user via a decision tree or a text-based explanation, or may not be shown at all. In this research, I conduct and design two human-subject studies; first, we conduct a study relating different abstractions of the cobot's policy to their induced situational awareness (SA) levels, measuring how different explanations can help a human perceive the current environment (Level 1), comprehend the AI's decision-making model (Level 2), and project into the future to develop a collaboration plan (Level 3). We find that cobots with xAI-based support can provide human teammates a higher level of SA, benefiting a human teammate's ability to perform situational analysis and understand the HMT scenario ($p < 0.05$). Second, we conduct a study on ad hoc human-machine teaming assessing how online xAI-based support, generated via cobot abstractions, and the human's ability to process higher levels of information affect teaming performance. We find that novices benefit from xAI-based support ($p < 0.05$) but are susceptible to information overload from more involved xAI abstractions ($p < 0.05$). Expert performance, on the other hand, degrades with the addition of xAI-based support ($p < 0.05$), indicating that the cost of paying attention to the explanation outweighs the benefits obtained from generating an accurate mental model of the cobot's behavior. We advance the aspiration of democratizing cobots by focusing on developing higher levels of understanding between humans and cobots.

Effective Communication with Humans

I have provided several contributions to allow cobots to better team with heterogeneous humans and provide a human teammate with an increased understanding of a cobot policy via interpretability. However, as prior work and our previous study displayed, humans have a limited cognitive bandwidth and explanations can degrade performance. Accordingly, our next step was to ensure that cobots are able to selectively and efficiently share information with humans. In high-performing human teams, human experts judiciously choose when to communicate and whom to communicate with, communicating only when beneficial (Salas, Cooke, and Rosen 2008). Each team member exhibits the role of a communicator and message receiver, relaying information to the right teammates and incorporating received information effectively. We would like to instill a similar behavior into how cobots communicate and coordinate with humans. In Niu, Paleja, and Gombolay (2021), we develop a cobot graph communication protocol, Multi-Agent Graph-attentIon Communication (MAGIC), that emulates the features of an effective human-human team by learning robot policies that determine "when" and "whom" with to communicate via an end-to-end framework, resulting in highly efficient communication among agents.

Future Work There are several items to consider in future work. Firstly, we would like to extend our model for heterogeneous LfD, the PNT, for continuous action spaces, allowing for an interpretable tree-based model for continuous control. Next, I would like to conduct studies about the effectiveness of cobots with interpretable policies and cobots with targeted communication for larger human-machine teams. Lastly, I would like to increase the applicability of my research by extending the ideas of heterogeneity across human teammates to multi-agent coordination with heterogeneous robots (robots with different actuators/sensors).

References

Chen, J. Y.; and Barnes, M. 2014. Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems*, 44: 13–29.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Niu, Y.; Paleja, R. R.; and Gombolay, M. 2021. Multi-Agent Graph-Attention Communication and Teaming. In *AAMAS*.

Paleja, R.; Ghuy, M.; Arachchige, N. R.; Jensen, R.; and Gombolay, M. 2021. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Paleja, R. R.; and Gombolay, M. 2019. Heterogeneous Learning from Demonstration. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 730–732.

Paleja, R. R.; Silva, A.; Chen, L.; and Gombolay, M. 2020. Interpretable and Personalized Apprenticeship Scheduling: Learning Interpretable Scheduling Policies from Heterogeneous User Demonstrations. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Salas, E.; Cooke, N.; and Rosen, M. 2008. On Teams, Teamwork, and Team Performance: Discoveries and Developments. *Human Factors: The Journal of Human Factors and Ergonomic Society*, 50: 540 – 547.

Sammut, C.; Hurst, S.; Kedzier, D.; and Michie, D. 2002. *Learning to Fly*, 171–189. Cambridge, MA, USA: MIT Press.