Autonomous Policy Explanations for Effective Human-Machine Teaming

Aaquib Tabrez

University of Colorado Boulder
Boulder, Colorado 80309
mohd.tabrez@colorado.edu

Abstract

Policy explanation, a process for describing the behavior of an autonomous system, plays a crucial role in effectively conveying an agent's decision-making rationale to human collaborators and is essential for safe real-world deployments. It becomes even more critical in effective human-robot teaming, where good communication allows teams to adapt and improvise successfully during uncertain situations by enabling value alignment within the teams. This thesis proposal focuses on improving human-machine teaming by developing novel human-centered explainable AI (xAI) techniques that empower autonomous agents to communicate their capabilities and limitations via multiple modalities, teach and influence human teammates' behavior as decision-support systems, and effectively build and manage trust in HRI systems.

Introduction and Research Themes

A shared understanding among teammates is crucial for effective teamwork; it helps them anticipate and align with each other's actions, leading to better decisions. While people are naturally good at this, robots are not. Previous research has demonstrated that explanations offer transparency and also play a functional role in synchronizing expectations during misalignments between human and robot teams (Chakraborti, Sreedharan, and Kambhampati 2020). Moreover, people tend to trust autonomous agents more when they have a clear understanding of the robot's capabilities and decision-making process.

Additionally, according to the "*ultra-strong*" criteria set by (Michie 1988), a machine learning system should not only be able to explain its hypothesis to a human but also be capable of teaching it, thereby enhancing the human's performance beyond just studying the data.

In my research I develop novel methodologies that enable these agents to effectively communicate and explain their decision-making rationale, as well as teach and enhance the understanding of collaborators to improve their behavior. Specifically, my research focuses on the following themes: 1) Operationalizing multimodal policy explanations for autonomous agents; 2) Characterizing a human-centered explainable robot coaching framework to enhance shared awareness; and 3) Evaluating the role of robot justification in mediating trust within human-machine teams.

Semantic Explanations for Robot Coaching

One of my goals is to use xAI to transform robots into effective coaches, ensuring value alignment among teammates. We introduced a coaching framework, Reward Augmentation and Repair through Explanation (RARE) (Tabrez, Agrawal, and Hayes 2019).

This framework encompasses: 1) inferring the human collaborator's task comprehension and estimating their reward function using Hidden Markov Models, 2) identifying missing components of the reward function with a Partially Observable Markov Decision Process, and 3) offering natural language explanations to address these misalignments. We conducted a study using a collaborative color-based sudoku game with an autonomous robotic arm to assess the RARE framework. The results indicated that when robots provided justifications for their actions (by explaining the potential mistakes users were about to make), they were perceived as more *helpful, useful, and intelligent*. In scenarios with justifications, irreversible mistakes dropped to 20%, compared to 80% without them – emphasizing that users trust robots more when they explain their corrective actions.

While the RARE effectively corrects a single instance of suboptimal human action, it is time-consuming and lacks context in its explanations. For instance, in an emergency evacuation where an autonomous agent is tasked with guiding people out of a building, a first-time visitor might not know how to react to a specific statement like "There's a fire near Conference Room 3." However, they could more easily adapt their plan if told, "The north half of the building is on fire" – highlighting the need for context-aware explanations.

Therefore, I am currently focused on creating a new optimization algorithm that uses semantic explanations, drawn from planning predicates, to improve agents' reward functions and behavior. While previous attempts to generate such explanations have been computationally expensive (Hayes and Shah 2017), our method employs a novel integer programming approach to efficiently solve the minimum set cover problem and adds policy elicitation to improve the collaborator's task performance. We have tested our algorithm's effectiveness in two real-world applications: robotic cleaning and emergency evacuation. Our method substan-

tially outperforms the previous best solution (Hayes and Shah 2017) to an extent that it is now practical for online use. We're in the process of conducting human subjects studies to assess the utility of our method's explanations.

Augmented-Reality for Visual Explanations

In situations with high uncertainty and continually evolving conditions, semantic explanations are not ideal. Visual information becomes more effective in such cases, especially when multiple likely hypotheses need to be portrayed as plans change based on new observations (i.e., partially observable domains). This inspired our work on AR-based visual guidance through a system called MARS (Min-entropy Algorithm for Robot-supplied Suggestions) (Tabrez, Luebbers, and Hayes 2022).

MARS is a multiagent reinforcement learning and planning algorithm, designed to address multi-goal tasks under uncertainty, that generates proactive visual recommendations. It uses a probability mass function (PMF) to represent uncertainty around whether a state is a goal, serving as a shared utility for both human and autonomous agents. The system employs online reinforcement learning to determine optimal policies for autonomous agents and action recommendations for human teammates. MARS also introduces two AR-based visual guidance types: prescriptive (visualizing recommended actions) and descriptive (visualizing state space information for decision-making).

We evaluated the MARS system in a human-subjects study using a 3D AR-based human-robot Minesweeper game. Participants experienced three conditions: prescriptive guidance, descriptive guidance, and combined guidance. Our findings support the hypothesis that combining environmental insight (descriptive guidance) with action suggestions (prescriptive guidance) enhances trust, interpretability, performance, and and made users more independent.

Autonomous Counterfactual Policy Justifications

In the MARS study, participants were frustrated by the system's unexpected behaviors, such as sudden path changes. This unpredictability stemmed from policy optimization in uncertain situations, leading to varied trust levels in the system; some participants over-trusted it while others under-trusted it. Participants perceived this emergent behavior as unconfident and expressed a desire for explanations, along with a mechanism to judge the quality of recommendations, echoing previous findings (Tabrez, Agrawal, and Hayes 2019). This motivated us in (Luebbers et al. 2023) to evaluate when justifications are most impactful and what information they should include to enhance human decision-making.

In this work, we developed a novel mathematical framework grounded in the value of information theory to identify the optimal timing for a robot to justify its recommendations to a human teammate. This framework was validated through an expert-feedback study, revealing that our strategic timing for justifications received the highest average rating for perceived usefulness compared to constant or timed-interval justifications.

We also introduced a methodological characterization of four distinct justification types: global policy, local policy, global environment, and local environment. These types were evaluated through an online human-subjects study. Our findings revealed that policy-based justifications promote higher compliance and quicker decision-making, while environment-based justifications enhance perceptions of a robot's interpretability, intelligence, and trustworthiness. Based on these insights, we recommended using policy-based justifications when the robot has high competence or the human has low competence. Conversely, environment-based justifications are best suited for situations with a less competent robot or a highly competent human.

Future Work

For future work, I plan to enhance my policy explanation techniques by integrating them with foundation models. The current policy explanation methods are sensitive to environmental changes and demand extensive domain-specific hand engineering. By leveraging foundation models, I hope to address these challenges and create more robust and generalizable systems suitable for real-world deployment.

In addition, I am interested in exploring the development of verifiable foundation models to improve human-machine communication and policy explanations. A significant challenge with foundation models is their propensity for hallucination, which can have catastrophic consequences, especially when deployed in safety-critical contexts where over-reliance on these systems is common.

Acknowledgments

This work was funded by the Army Research Lab STRONG Program (#W911NF-20-2-0083).

References

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2020. The emerging landscape of explainable ai planning and decision making. *arXiv preprint arXiv:2002.11697*.

Hayes, B.; and Shah, J. A. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 303–312. IEEE.

Luebbers, M.; Tabrez, A.; Ruvane, K.; and Hayes, B. 2023. Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. In *Proceedings of Robotics: Science and Systems*. Daegu, Republic of Korea.

Michie, D. 1988. Machine learning in the next five years. In *Proceedings of the 3rd European conference on European working session on learning*, 107–122.

Tabrez, A.; Agrawal, S.; and Hayes, B. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 249–257. IEEE.

Tabrez, A.; Luebbers, M. B.; and Hayes, B. 2022. Descriptive and prescriptive visual guidance to improve shared situational awareness in human-robot teaming. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1256–1264.