Mediating Trust and Influence in Human-Robot Interaction via Explainable AI

Aaquib Tabrez
University of Colorado Boulder, Colorado, USA
mohd.tabrez@colorado.edu

Bradley Hayes
University of Colorado Boulder, Colorado, USA
bradley.hayes@colorado.edu

I. Introduction and Motivation

For robots to effectively collaborate with humans in high-stakes applications (e.g., autonomous driving), insights into these autonomous systems' capabilities and their limitations are required [26, 18, 21]. Our work leverages explainable AI (xAI) techniques to provide those insights, enabling more fluent teaming and agent-to-human communication. While most recent literature in robotics focuses on enabling robots to adapt to their human teammates (e.g., imitation learning) [1, 19], in this work, we focus on the converse, empowering autonomous agents to be capable of manipulating and adapting their teammates' behavior during joint task execution.

One critical aspect for safe and effective collaboration between teammates is maintaining awareness over the collaborator's mental model, enabling agents to reason about what their teammate is likely to do or need [6]. While people are quite skillful in this task, robots lack this intuition and capability. As described in our survey on mental modeling techniques in human-robot teaming [22], researchers have leveraged xAI for knowledge sharing and expectation matching to achieve fluent collaboration and improve shared awareness [3]. Explanations enhance transparency and functionally help in the synchronization of expectation between the human and robot teams [2].

With this in mind, we pursue two research themes at the intersection of xAI and human-robot interaction: **RT1:** Formulate and operationalize a framework for explainable robot coaching within human-robot teaming scenarios to improve shared awareness, **RT2:** Characterize and generate semantic and visual modalities for robot explanation, and evaluate the role of robot justification on mediating trust and eliciting desired behavior within human-machine teams.

II. Prior Work

A. Semantic Explanations and Justifications

Framework for Robot Coaching and Justification. One of our goals is to transform robots into competent coaches, using explainable AI to establish shared mental models amongst teammates. Therefore, we developed a novel robot coaching framework called Reward Augmentation and Repair through Explanation (RARE) [20]. The core functionality is as follows: 1) RARE infers the collaborator's task understanding, estimating their reward function using Hidden Markov Models, 2) it identifies missing components of the reward function via a Partially Observable Markov Decision Process, and 3) it provides natural language explanations to facilitate reward function repair, improving task comprehension.

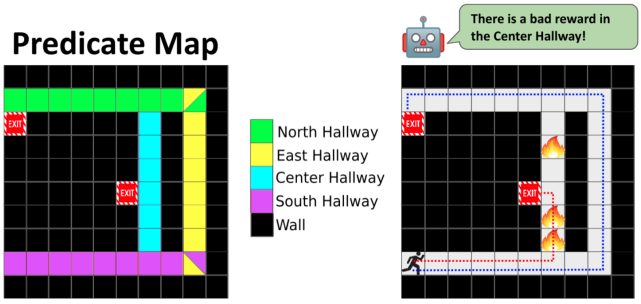

Fig. 1. A human agent attempts to exit the building in an emergency evacuation scenario (right), lacking knowledge about the fires. SPEAR leverages semantic updates using predicates (left) to produce optimal behavior.

Through a between-subjects user study, we evaluated the viability and effectiveness of RARE using a collaborative color-based sudoku game, where users teamed with an autonomous robotic arm. The experiment compared two study conditions varying by the content provided during a robot interruption. The control consisted of a simple indication that the user is about to make a mistake leading to task failure, while the justification condition contained additional information explaining the reason for the future failure.

We found statistically significant support across subjective measures to validate the hypothesis that participants found robots more helpful, useful, and intelligent when they provide justifications. Objectively, we observed more game terminations (irreversible mistakes) during the control condition than the justification condition (80% vs. 20%). Our exit survey showed that people did not trust the robot when it intervened without further explanation (e.g., the reason for game termination), indicating justification is likely necessary when a robot corrects users or recommends alternate actions.

One-shot Policy Elicitation via Semantic Explanations. RARE corrects a single instance of suboptimal human action at a time, which can be tedious and time-consuming for human collaboration. Furthermore, RARE does not consider the recipient's world model, leading to the generation of uninterpretable explanations. Consider an emergency evacuation scenario, where an agent is tasked with guiding people safely out of a building. Someone visiting for the first time may not

know how to change their evacuation plan when told, "There's a fire near Conference Room 3", but may be able to adapt their plan if told "the north half of the building is on fire".

Thus, we proposed Single-shot Policy Explanation for Augmenting Rewards (SPEAR) [23], a novel optimization algorithm that uses semantic explanations derived from combinations of planning predicates to augment agents' reward functions, driving their policies to exhibit more optimal behavior. Predicates are pre-defined boolean state classifiers (as found in traditional STRIPS planning [9]) with associated string explanations, as shown in Figure 1-left. Prior work solves natural language generation as a set cover problem to find the smallest logical expression of predicates, but their solution is of exponential runtime, preventing its use in most real-world problems [11]. We solve the minimum set cover using a novel integer programming formulation and policy elicitation to improve the collaborator's task performance (Figure 1).

We experimentally validated our algorithm's policy elicitation capabilities in two practically grounded applications: 1) a robotic cleaning task, and 2) an emergency evacuation scenario. The first scenario validated the live deployment of SPEAR within two robotic agents, where one agent needs to correct the policy of a second robotic agent to prevent it from removing specific items during a pick and place task. The second task analyzed the performance of SPEAR on both stochastic and deterministic domains (Figure 1). Our approach outperforms prior work [11] by multiple orders of magnitude.

B. Visual Explanations

Descriptive and Prescriptive Visual Guidance.

Semantic explanations are not well suited for certain scenarios, especially those involving high uncertainty, requiring the portrayal of multiple competent hypotheses as plans change based on new observed information (i.e., partially observable domains). For these continually evolving domains, visual information representation is ideal [8], motivating our subsequent work on AR-based visual guidance called MARS (Min-entropy Algorithm for Robot-supplied Suggestions) [24].

MARS consists of a planning algorithm for uncertain environments, informing the generation of proactive visual recommendations. Environmental uncertainty is characterized as a dynamically-updating probability mass function (PMF), a common practice across various classes of search task [12, 5, 13]. The PMF serves as a shared utility function common to all agents (both human and autonomous), providing insight into the agent's policy. This PMF is utilized by two separate Markov Decision Processes (MDPs); one for autonomous agents, and another for generating assistive guidance for the human teammate. MARS solves both of these MDPs via online reinforcement learning to get optimal policies for autonomous agents and action recommendations for human teammates respectively. We also provided a characterization of two distinct AR-based visual guidance modalities: prescriptive guidance (visualizing recommended actions) and descriptive guidance (visualizing state space information to aid in decision-making), as shown in Figure 2.

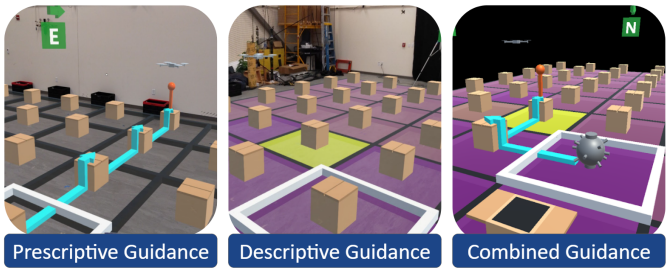

Fig. 2. AR-based visual guidance: prescriptive guidance - arrows and pins (left), descriptive guidance - an environmental heatmap (middle), and a combination of both (right) in an Minesweeper-inspired domain

We evaluated the utility of our visual guidance modalities and the effectiveness of the MARS algorithm through a within-subjects human study using a human-robot collaborative analogue of the PC game Minesweeper, played using a HoloLens 2 AR headset. Participants experienced three conditions based on the type of visual guidance given to the human teammate as informed by sensor readings from a virtual drone: 1) prescriptive guidance, 2) descriptive guidance, and 3) a combination of prescriptive and descriptive guidance (Figure 2). We found statistical significance supporting our hypothesis that combining visual insight into environmental uncertainty (descriptive guidance) with robot-provided action suggestions (prescriptive guidance) improved trust, interpretability, and performance, and made human collaborators more independent.

III. Future Work

In the MARS study, some participants disliked when the system's recommendation exhibited unexpected behavior (e.g., a sudden path change). These inexplicable recommendations resulted from policy optimization within an uncertain environment. People viewed this emergent behavior as confusing and unconfident, expressing the desire to receive explanations when this happens, echoing previous findings [7]. Our subsequent research will attempt to address this challenge via visual counterfactual justifications. Counterfactuals are an xAI technique demonstrating how specific changes to the inputs of known models would lead to output classification changes [14, 17]. These explanations can provide context about internal model reasoning to users, leading to usefulness for model debugging and failure recovery [16, 10, 4, 25].

Similarly, we noticed some people in the study over-trusted the guidance (taking its suggestions to be inherently correct), while others under-trusted it (frequently ignoring good advice). The exit interviews indicated that people did not have an appropriate idea of judging the quality of recommendations, leading to variable perceived system reliability. Therefore, we are working on different formulations of justification to help users appropriately assess robot recommendations for mitigating over- and under-trust (e.g., semantically indicating likelihood) [15]. Simultaneously, we are developing an algorithmic framework that can incorporate these explanations, leveraging SPEAR and MARS within uncertain domains to improve reliance and trust in human-robot teaming scenarios.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57 (5):469–483, 2009.

[2] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The emerging landscape of explainable automated planning & decision making.

[3] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*, 2017.

[4] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020.

[5] THJ Collett and Bruce A MacDonald. Developer oriented visualisation of a robot program. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 49–56, 2006.

[6] Nancy J Cooke, Eduardo Salas, Janis A Cannon-Bowers, and Renee J Stout. Measuring team knowledge. *Human factors*, 42(1):151–173, 2000.

[7] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, pages 507–513, 2018.

[8] Bruce H Deatherage. Auditory and other sensory forms of information presentation. *Human engineering guide to equipment design*, pages 123–160, 1972.

[9] Richard E Fikes and Nils J Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208, 1971.

[10] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

[11] Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 303–312. IEEE, 2017.

[12] Haikun Huang, Ni-Ching Lin, Lorenzo Barrett, Darian Springer, Hsueh-Cheng Wang, Marc Pomplun, and Lap-Fai Yu. Automatic optimization of wayfinding design. *IEEE transactions on visualization and computer graphics*, 24(9):2516–2530, 2017.

[13] Tijn Kooijmans, Takayuki Kanda, Christoph Bartneck, Hiroshi Ishiguro, and Norihiro Hagita. Interaction debugging: an integral approach to analyze human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 64–71, 2006.

[14] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[15] Matthew B Luebbers, Aaquib Tabrez, and Bradley Hayes. Augmented reality-based explainable ai strategies for establishing appropriate reliance and trust in human-robot teaming. *5th International Workshop on Virtual, Augmented and Mixed Reality for HRI (VAM-HRI)*, 2022.

[16] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Distal explanations for model-free explainable reinforcement learning. *arXiv preprint arXiv:2001.10284*, 2020.

[17] Tim Miller. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163*, 2018.

[18] Aaquib Tabrez and Bradley Hayes. Improving human-robot interaction through explainable reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 751–753. IEEE, 2019.

[19] Aaquib Tabrez, Jack Kawell, and Bradley Hayes. Asking the right questions: Facilitating semantic constraint specification for robot skill learning and repair. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6217–6224. IEEE.

[20] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 249–257. IEEE, 2019.

[21] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. Automated failure-mode clustering and labeling for informed car-to-driver handover in autonomous vehicles. *arXiv preprint arXiv:2005.04439*, 2020.

[22] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1(4):259–267, 2020.

[23] Aaquib Tabrez, Ryan Leonard, and Bradley Hayes. One-shot policy elicitation via semantic reward manipulation. *arXiv preprint arXiv:2101.01860*, 2021.

[24] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. 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*, 2022.

[25] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1784–1793, 2021.

[26] Adriana Tapus, Maja J Mataric, and Brian Scassellati. Socially assistive robotics [grand challenges of robotics]. *IEEE robotics & automation magazine*, 14(1):35–42, 2007.