Effective Human-Machine Teaming through Communicative Autonomous Agents that Explain, Coach, and Convince

Doctoral Consortium

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

ABSTRACT

Effective communication is essential for human-robot collaboration to improve task efficiency, fluency, and safety. Good communication between teammates provides shared situational awareness, allowing them to adapt and improvise successfully during uncertain situations, and helps identify and remedy any potential misunderstandings in the case of incongruous mental models. This doctoral proposal focuses on improving human-agent communication by leveraging explainable AI techniques to empower autonomous agents to 1) communicate insights into their capabilities and limitations to a human collaborator, 2) coach and influence human teammates' behavior during joint task execution, and 3) successfully convince and mediate trust in human-robot interactions.

KEYWORDS

Human-agent Collaboration; Explainable AI; Shared Mental Models; Reinforcement Learning; Augmented Reality; Policy Explanations

ACM Reference Format:
Aaquib Tabrez. 2023. Effective Human-Machine Teaming through Communicative Autonomous Agents that Explain, Coach, and Convince: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION AND RESEARCH THEMES

Having a shared understanding of the task and the environment is crucial for safe and efficient collaboration among team members. Shared mental models allow agents to anticipate the actions and needs of their teammates, which enables them to coordinate their actions and make better decisions [5]. While people are quite skillful in this task, robots lack this intuition and capability. As described in our survey [21], researchers have leveraged explainable AI (xAI) for knowledge sharing and expectation matching to achieve fluent collaboration and improve shared awareness [2, 4, 14, 19, 23].

Explanations enhance transparency and functionally help synchronize expectations when there is an incongruity between human and robot teams [3, 17]. Moreover, people trust autonomous agents more when they understand the agents' roles and responsibilities, have confidence in their abilities, and possess a clear understanding of their decision-making processes [1, 15]. Therefore, it is essential to develop new methodologies that enable these agents to effectively communicate and explain their decision-making rationale, thereby gaining the trust of their human teammates [12, 16].

Our research focuses on three interconnected themes:
RT1: Characterizing and generating explanations for autonomous agents to effectively communicate their decision-making rationales, **RT2:** Operationalizing a framework for explainable robot coaching within human-robot teaming scenarios to improve shared awareness, **RT3:** Evaluating the role of robot justification in mediating trust and eliciting desired behavior within human-machine teams.

2 PRIOR WORK

2.1 Semantic Explanations

Framework for Robot Coaching and Justification. One of our objectives is to transform robots into competent coaches by utilizing xAI to establish shared mental models among teammates. We developed a novel robot coaching framework called Reward Augmentation and Repair through Explanation (RARE) [18]. The central functionality of RARE is comprised of the following steps: 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 via a Partially Observable Markov Decision Process, and 3) generating and providing natural language explanations to facilitate reward function repair.

We evaluated the feasibility and efficiency of RARE through a between-subjects user study, using a collaborative color-based sudoku game, where users worked with an autonomous robotic arm. Our study found evidence to support the hypothesis that providing justifications can improve users' perceptions of robots. The study compared two conditions, varying by the content provided during a robot interruption. The control condition consisted of a simple indication that the user was about to make a mistake leading to task failure, while the justification condition included additional information explaining the reason for the future failure.

Subjective measures showed that participants found the robot more helpful, useful, and intelligent when justifications were provided. Objective measures also revealed that there were fewer irreversible mistakes in the justification condition (20%) compared to the control condition (80%). Our exit survey results further highlighted that people were less likely to trust the robot when it intervened without providing explanations, emphasizing the importance of justification when robots correct users.

One-shot Policy Elicitation via Semantic Explanations. While the RARE framework effectively corrects a single instance of suboptimal human action, it can be tedious and time-consuming for human

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

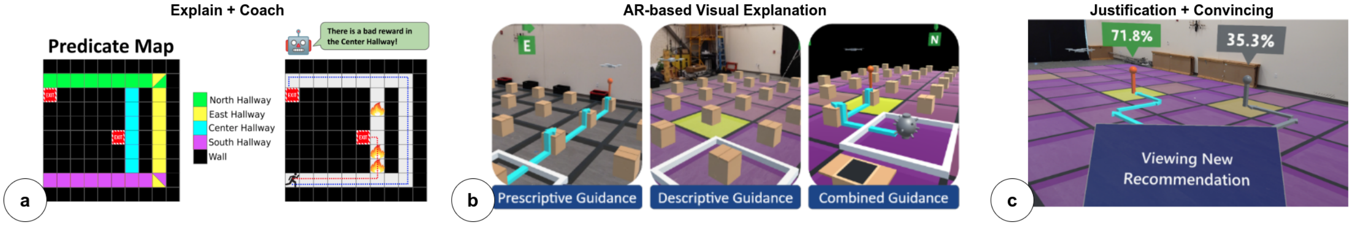

Figure 1: Components of human-machine communication: a) Explain + Coach: A human agent attempts to exit the building in an emergency evacuation scenario (right), lacking knowledge about the fires. SPEAR leverages semantic updates using predicates (left) to produce optimal behavior. b) Visualize + Recommend: AR-based visual explanations: prescriptive guidance - arrows and pins (left), descriptive guidance - an environmental heatmap (middle), and a combination of both (right) in a Minesweeper-inspired domain. c) Justification + Convincing: Prototype counterfactual alert in the Minesweeper domain.

collaboration. Additionally, RARE does not take into account the recipient's world model, resulting in the generation of explanations that may not be easily actionable. For example, in an emergency evacuation scenario, where an autonomous agent is tasked with guiding people safely out of a building, someone visiting the building 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.” This highlights the importance of considering the recipient's world model and providing context-specific explanations.

Thus, we proposed Single-shot Policy Explanation for Augmenting Rewards (SPEAR) [20], a novel optimization algorithm that utilizes semantic explanations derived from combinations of planning predicates to augment agents' reward functions and improve their behavior. Predicates are pre-defined Boolean state classifiers (as found in traditional STRIPS planning [8]) with associated string explanations (Figure 1a-left). Previous work has attempted to generate natural language using a set cover problem, but their solution has exponential runtime, preventing its use in most real-world problems [10]. Our approach solves the minimum set cover using a novel integer programming formulation and adds policy elicitation to improve the collaborator's task performance (Figure 1a).

We experimentally validated the capabilities of our algorithm in two practical applications: 1) a robotic cleaning task, and 2) an emergency evacuation scenario. Our approach outperformed the prior state of the art [10] by multiple orders of magnitude.

2.2 Augmented-Reality for Visual Explanations

Descriptive and Prescriptive Visual Explanations. Semantic explanations are not ideal for certain scenarios, particularly those involving high uncertainty, where multiple competent hypotheses need to be portrayed as plans change based on new observations (i.e., partially observable domains). In these continually evolving domains, visual information presentation is more effective [7]. This motivated our subsequent work on AR-based visual guidance called MARS (Min-entropy Algorithm for Robot-supplied Suggestions) [22].

MARS consists of a planning algorithm for uncertain environments, informing the generation of proactive visual recommendations. Environmental uncertainty is represented by a probability mass function (PMF) that serves as a shared utility function for all agents (both human and autonomous), providing insight into the agent's policy. MARS uses online reinforcement learning to find optimal policies for autonomous agents and action recommendations for human teammates. We also classified two AR-based visual guidance modalities: prescriptive guidance (recommended actions visualization) and descriptive guidance (state space information visualization to support decision-making), shown in Figure 1b.

We evaluated the effectiveness of our visual guidance modalities and the MARS algorithm through a within-subjects study using a 3D AR-based human-robot collaborative analog of the PC game Minesweeper. Participants experienced three conditions based on the type of guidance given to the human teammate as informed by sensor readings from a virtual drone: 1) prescriptive guidance, 2) descriptive guidance, and 3) both prescriptive and descriptive guidance (Figure 1b). 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 users more independent.

3 FUTURE WORK

In the MARS study, some participants were frustrated when the system's recommendations exhibited unexpected behavior, such as sudden path changes. These inexplicable recommendations resulted from policy optimization within an uncertain environment. Participants viewed this emergent behavior as confusing and unconfident, expressing a desire for explanations, echoing previous findings [6]. Similarly, we noticed that some participants 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 participants did not have an appropriate way of judging the quality of recommendations, leading to variable perceived system reliability.

To address these challenges, we are developing multiple formulations of justifications and policy visualization using counterfactual explanations to help users appropriately assess an agent's decision-making rationale and mitigate over- and under-trust, as shown in Fig 1c [12]. Counterfactual explanations show how changing inputs affects output classification and aid in providing context to users, model debugging, and failure recovery [9, 13]. Simultaneously, we are developing a formal framework using value of information theory [11] to strategically time justifications during periods of misaligned expectations for greater effect while improving compliance and trust within human-agent teaming scenarios.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[2] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan explanations as model reconciliation–an empirical study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 258–266.

[3] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. [n.d.]. The Emerging Landscape of Explainable Automated Planning & Decision Making.

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

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

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

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

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

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

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

[11] Tobias Kaupp, Alexei Makarenko, and Hugh Durrant-Whyte. 2010. Human–robot communication for collaborative decision making—A probabilistic approach. *Robotics and Autonomous Systems* 58, 5 (2010), 444–456.

[12] Matthew B Luebbers, Aaquib Tabrez, and Bradley Hayes. 2022. 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).

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

[14] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[15] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[16] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

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

[19] Aaquib Tabrez, Jack Kawell, and Bradley Hayes. [n.d.]. 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)*. IEEE, 6217–6224.

[20] Aaquib Tabrez, Ryan Leonard, and Bradley Hayes. 2021. One-shot Policy Elicitation via Semantic Reward Manipulation. *arXiv preprint arXiv:2101.01860* (2021).

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

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

[23] Luca Vigano and Daniele Magazzeni. 2020. Explainable security. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 293–300.