Explainable Guidance and Justification for Mental Model Alignment in Human-Robot Teams

Matthew B. Luebbers
matthew.luebbers@colorado.edu
University of Colorado Boulder
Boulder, CO, USA

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

ABSTRACT

There is potential for humans and autonomous robots to perform tasks collaboratively as teammates, achieving greater performance than either could on their own. Productive teamwork, however, requires a great deal of coordination, with human and robot agents maintaining well-aligned mental models regarding the shared task and each agent's role within it. Achieving this requires live and effective communication, especially as plans change due to shifts in environment knowledge. Our work leverages augmented reality and natural language interfaces to recommend policies to human teammates, explain the rationale of those policies, and justify during times of mismatched expectation, facilitating plan synchronization in partially observable, collaborative human-robot domains.

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**;
• **Computer systems organization** → **Robotics**.

KEYWORDS

human-robot teaming, augmented reality, mental models

ACM Reference Format:
Matthew B. Luebbers and Bradley Hayes. 2024. Explainable Guidance and Justification for Mental Model Alignment in Human-Robot Teams. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3610978.3638364

1 INTRODUCTION AND RELATED WORK

Robotic deployments have traditionally fallen into one of two distinct paradigms: autonomy or teleoperation. However, a third paradigm has received much attention in the past twenty years: humans and robots working together as teammates [6, 14, 15]. The key insight behind this operational model is that humans and autonomous systems excel at different things [7]. By effectively leveraging humans' and robots' specialized capabilities for joint tasks, team performance can exceed the mere sum of its parts.

Integrating a human into a multi-agent planner is incredibly difficult, however, since the inherent uncertainty of human behavior hinders effective optimization. To achieve nominal performance, let alone performance gains, human and robot teammates must synchronize their planning through a shared mental model [19]. To achieve this, effective communication between agents is required.

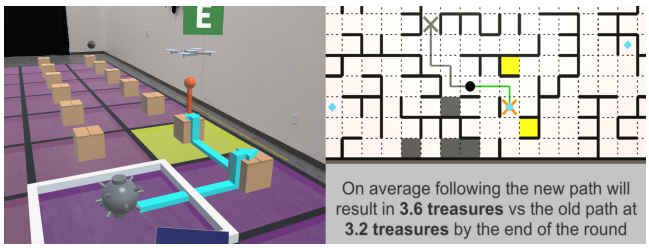

Figure 1: Left: visual guidance generated by the MARS algorithm, shown in AR as a combination of action recommendations (arrows and pin) and environmental probability data (heatmap). Right: a justification generated by our framework, providing rationale for why a human should follow new guidance (green arrow) over prior guidance (gray arrow).

One technique we leverage is augmented reality (AR) visualization, a technology whose capabilities have already been demonstrated in multiple robotic domains [4, 16, 21], including works in which we ourselves have shown AR's ability to facilitate smooth human-robot coordination in tabletop manipulation environments [12] and shared warehouse floors [5]. AR possesses the unique ability to project data directly onto the environment. This in-situ visualization gives shared environmental context for the human and the robot, enabling compact visual communication without the need for context switching to a separate screen [8, 10].

We also take inspiration from explainable AI, which has been shown not only to increase understanding of opaque learning models [1, 9], but also to promote team fluency and improve shared awareness in human-robot tasks [2, 3, 18]. In our work, we use algorithmically-backed AR visualization and natural language explanation to serve as the communicative bridge necessary to integrate humans into multi-agent reinforcement learning (RL) planners, solving collaborative, partially-observable tasks by leveraging each agent's unique skillset. This abstract describes works addressing two research questions: **Q1**. how should robots communicate to humans while performing tasks under uncertainty to enhance team performance? And **Q2**. how can robots justify their decisions and guidance to human teammates to improve trust and compliance?

2 EXPLAINABLE DECISION SUPPORT: MARS

We motivate this work with a prototypical search and rescue domain. If autonomous aerial drones equipped with sensors could be deployed to assist ground teams in sweeping the environment,

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3638364

rescue efforts could be greatly expedited, provided the team was supplied with the right inter-agent communication. To address this, we developed MARS (Min-Entropy Algorithm for Robot-Supplied Suggestions) [20], a multi-agent collaborative planning algorithm which simultaneously controls robot teammates and generates proactive, AR-based visual recommendations for human teammates.

MARS characterizes uncertainty about the location of goals in an environment through the use of a dynamically updating probability mass function (PMF) indicating the likelihood of a goal at each state. This PMF is incorporated into the reward signal for parallel Markov Decision Processes, one generating autonomous agent policies and one generating human recommendations, capturing the difference in objectives and capabilities for each agent class.

The human recommendations are delivered via AR headset and come in two flavors: prescriptive (directly recommending actions in the form of arrows showing where the system thinks the humans should go), and descriptive (displaying environmental data that informed the recommendation in the form of an evolving PMF heatmap, using a color gradient to represent likelihood of finding a target at a given location, from purple (low probability) to yellow (high): see Fig. 1 left). While prescriptive guidance is easy to understand and to follow, it leaves its decision-making process opaque. Descriptive guidance, on the other hand, allows users to consider all available information at the cost of extra cognitive workload.

To evaluate and compare these guidance modalities, we ran a user study, where participants played a collaborative 3D AR-based game inspired by Minesweeper in a large space alongside a virtual drone teammate. The goal of the game was to locate and defuse all mines hidden throughout the environment as quickly as possible. Guidance was provided by the drone teammate, using a noisy mine-detecting sensor to inform its recommendations. Between conditions, we varied what type of guidance we provided to the participant: prescriptive, descriptive, or both (Fig. 1 left).

The combined guidance scored highest on subjective measures of trust and interpretability, as well as objective task performance. What's more, participants were able to act with more independence using the combined guidance, deviating from drone-provided suggestions without degrading their performance. Participants used different strategies and thought patterns in the presence of different guidance types, following prescriptive guidance automatically while stopping to think carefully about their next move when given descriptive guidance. A combination of both guidance types led to the best performance, decreasing mental load by providing suggestions, but allowing humans to deviate strategically when necessary.

3 STRATEGIC POLICY JUSTIFICATION

Although prior work in explainable AI has shown the benefits of providing explanations to illuminate opaque systems [9], very little research has focused on the timing of such explanations. During the MARS study, whenever new drone observations caused a change in guidance, participants often became confused and frustrated, undermining their trust in and compliance with that guidance. In post-experiment surveys, many expressed a desire for explanation during these unexpected path changes. This inspired our next work [13], which aimed to answer both when explanations are appropriate or useful, and what content those explanations should include.

In this work, we were interested in leveraging explanation to serve as justification. We defined a **justification** as an explanation of an action or suggestion, timed strategically to align with a mismatch in expectation between agents. We developed a novel mathematical framework utilizing value of information (VOI) theory [11] to trigger justifications whenever human and robot policies diverge enough that the expected benefit to a human teammate of receiving an update exceeds the added workload of attending to it. We validated our VOI framework through a user study where participants viewed videos of human-agent tasks with justifications presented using varying timing strategies. We found that our VOI approach was rated as significantly more useful for decision-making compared with constant or timed-interval justifications.

We added a justification module using this VOI trigger to the MARS framework described earlier in [20]. To generate the justification content, we developed a dual-axis characterization of justification types, each rooted in an aspect of RL problems [17]. The first axis is justification basis: environment-based (relating to the environmental features that affected the change in poilcy) or policy-based (relating to the reward outcomes of the policy change). The second axis is justification scope: local (grounded in short-horizon contexts and subgoals) or global (grounded in the full task).

We evaluated four justification types: *global policy* (Fig. 1 right), *local policy*, *global environment*, and *local environment*, alongside a control condition with no justification, in an online human-subjects study involving a collaborative partially-observable treasure search task with guidance provided by a fleet of drones carrying noisy treasure-detection sensors. We found an interesting dichotomy in our results: while the policy-based conditions led to the highest compliance with guidance, fastest decision making, and best game performance, the environment-based justifications were consistently rated as more interpretable, trustworthy, and intelligent.

In our domain, compliance with guidance was highly correlated with performance, but this is not true of every domain. Therefore, we recommend using policy-based justifications in domains or situations where robots are likely to have high competence, eliciting swift compliance with guidance, and using environment-based justifications where robots are likely to have low competence, so users are nudged into more effortful and deliberate thought patterns.

4 IMPACT AND FUTURE WORK

These works have shown the ways in which live communication, leveraging explainable guidance and justification, can influence the performance of human-robot teams. Both works introduce novel algorithmic contributions and interfaces that allow humans to act within a multi-agent RL framework to solve collaborative, partially observable domains. The results of the user studies inform a number of takeaways for designing and deploying such systems in the field.

Our ongoing research directly builds on our MARS framework, first by expanding the algorithm's usefulness to a much wider array of complex use-cases through a novel, recursive spatial hierarchy technique. We are also exploring how differential guidance presented to human teammates in multi-human, multi-robot teams can influence the evolution of dynamics between human teammates, such as leader and follower roles, leading to more fluent teamwork and efficient mental load sharing among agents.

REFERENCES

[1] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing & Management* 60, 1 (2023), 103111.

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

[4] Kishan Chandan, Vidisha Kudalkar, Xiang Li, and Shiqi Zhang. 2019. Negotiation-based human-robot collaboration via augmented reality. *arXiv preprint arXiv:1909.11227* (2019).

[5] Christine T Chang, Matthew B Luebbers, Mitchell Hebert, and Bradley Hayes. 2023. Human Non-Compliance with Robot Spatial Ownership Communicated via Augmented Reality: Implications for Human-Robot Teaming Safety. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9785–9792.

[6] Anca D Dragan and Siddhartha S Srinivasa. 2012. *Formalizing assistive teleoperation*. MIT Press, July.

[7] Paul M Fitts. 1951. Human engineering for an effective air-navigation and traffic-control system. (1951).

[8] Scott A Green, Mark Billinghurst, XiaoQi Chen, and J Geoffrey Chase. 2008. Human-robot collaboration: A literature review and augmented reality approach in design. *International journal of advanced robotic systems* 5, 1 (2008), 1.

[9] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science robotics* 4, 37 (2019), eaay7120.

[10] Hooman Hedayati, Michael Walker, and Daniel Szafir. 2018. Improving collocated robot teleoperation with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 78–86.

[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, Connor Brooks, Carl L Mueller, Daniel Szafir, and Bradley Hayes. 2021. Arc-lfd: Using augmented reality for interactive long-term robot skill maintenance via constrained learning from demonstration. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3794–3800.

[13] Matthew B Luebbers, Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. 2023. Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. In *Proceedings of Robotics: Science and Systems*. Daegu, Republic of Korea. https://doi.org/10.15607/RSS.2023.XIX.002

[14] Stefanos Nikolaidis, Przemyslaw Lasota, Gregory Rossano, Carlos Martinez, Thomas Fuhlbrigge, and Julie Shah. 2013. Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action. In *IEEE ISR 2013*. IEEE, 1–6.

[15] Illah R Nourbakhsh, Katia Sycara, Mary Koes, Mark Yong, Michael Lewis, and Steve Burion. 2005. Human-robot teaming for search and rescue. *IEEE Pervasive Computing* 4, 1 (2005), 72–79.

[16] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. 2019. Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays. *The International Journal of Robotics Research* 38, 12-13 (2019), 1513–1526.

[17] Lindsay Sanneman and Julie A Shah. 2022. An empirical study of reward explanations with human-robot interaction applications. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8956–8963.

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

[20] 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*. 1256–1264.

[21] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. 2018. Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 316–324.