Explanation-based Reward Coaching to Improve Human Performance via Reinforcement Learning

Aaquib Tabrez
University Of Colorado Boulder
Boulder, CO 80309
mohd.tabrez@colorado.edu

Shivendra Agrawal
University Of Colorado Boulder
Boulder, CO 80309
shivendra.agrawal@colorado.edu

Bradley Hayes
University Of Colorado Boulder
Boulder, CO 80309
bradley.hayes@colorado.edu

***Abstract*—For robots to effectively collaborate with humans, it is critical to establish a shared mental model amongst teammates. In the case of incongruous models, catastrophic failures may occur unless mitigating steps are taken. To identify and remedy these potential issues, we propose a novel mechanism for enabling an autonomous system to detect model disparity between itself and a human collaborator, infer the source of the disagreement within the model, evaluate potential consequences of this error, and finally, provide human-interpretable feedback to encourage model correction. This process effectively enables a robot to provide a human with a policy update based on perceived model disparity, reducing the likelihood of costly or dangerous failures during joint task execution. This paper makes two contributions at the intersection of explainable AI (xAI) and human-robot collaboration: 1) The Reward Augmentation and Repair through Explanation (RARE) framework for estimating task understanding and 2) A human subjects study illustrating the effectiveness of reward augmentation-based policy repair in a complex collaborative task.**

***Index Terms*—Explainable AI; Policy Explanation; Human-Robot Collaboration; Reward Estimation; Joint Task Execution**

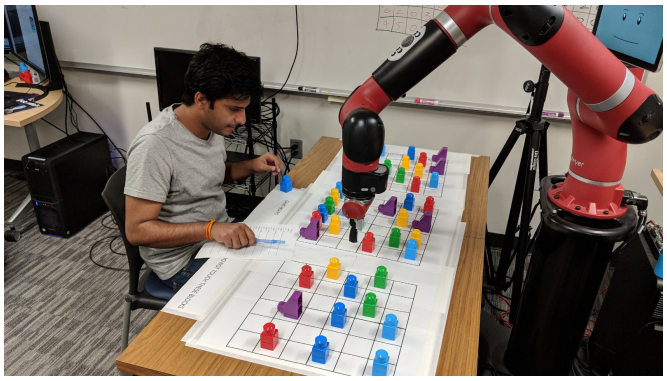


Fig. 1: A participant plays a collaborative, color-based Sudoku variant with a robot during a human subjects study evaluating the proposed framework. Using RARE, the robot is able to identify, indicate, and explain potential failure modes of the game based on the human's predicted understanding of the game's reward function.

I. Introduction

Shared expectations are crucial for fluent and safe teamwork. Establishing a common mental model of a task is essential for human-robot collaboration, where each team member's skills and knowledge may be combined to accomplish more than either could in isolation [16], [24], [30]. However, gaining insight into a collaborator's decision-making process during task execution can be prohibitively difficult, requiring the agent to have the capability to perform policy explanation [17]. Further, taking corrective actions when a team member's comprehension of the task doesn't match your own requires one to not just indicate a problem with the policy, but also to identify the root cause of the incongruousness.

Within society, the roles and responsibilities being assigned to robots have grown increasingly complex, reaching the boundaries of social integration. As this continues, it is reasonable to assume people will increasingly turn towards robots for completing important collaborative tasks with real consequences of failure, such as search and rescue [7], housekeeping [13], and personal assistance for the elderly [29], [33]. Providing these autonomous systems with the ability to identify and explain potential failures or root causes of sub-optimal behavior during collaboration will be essential to establishing appropriate levels of trust and reliance, while simultaneously improving the task understanding and performance of human operators.

Consider the problem of resource allocation and asset tasking during a collaborative search and rescue operation, where a human operator is commanding a fleet of UAVs. If the human provides a sub-optimal flight plan to an agent that provides poor coverage or exceeds its flight range, a system that could both generate human-interpretable feedback indicating the potential failure mode associated with the human's action and provide a justifying explanation would be far more useful than one that could not. One might expect such a capability to improve both operator task proficiency and failure rates.

To provide usable feedback for avoiding sub-optimal behaviors expected of a collaborator, we introduce a framework that leverages the assumption that sub-optimal collaborator behavior is the result of a misinformed understanding of the task rather than a problem with the collaborator's rationality. In terms of a task defined through a Markov Decision Process, a human's poor action selections should be attributable to a malformed reward function rather than a malformed policy search algorithm. Building on this assumption, we believe a useful autonomous collaborator should be able to 1) infer the most likely reward function used as a basis for a human's

behaviors; 2) identify the single most detrimental missing piece of the reward function; and 3) communicate this back to the human as actionable information.

Toward this goal, we propose *Reward Augmentation and Repair through Explanation* (RARE), a novel framework for improving human-robot collaboration through reward coaching. RARE enables a robot to perform policy estimation during a collaborative task and offer corrections to a teammate's mental model during joint task execution. Our model estimates the most likely reward function that explains the collaborator's behavior and provides a repairing explanation meant to enable the collaborator to update their reward function (task comprehension) and policy (behavior). The two primary contributions of our work are:

- Reward Augmentation and Repair through Explanation (RARE), a novel framework for understanding and correcting an agent's decision-making process, which estimates an agent's understanding of a domain's reward function through their behavior and provides corrective explanations to repair detected issues.
- A human subjects study-based evaluation of RARE, showing both the technical feasibility of the approach alongside empirical results illustrating its effectiveness during a complex human-robot collaboration.

II. Background and Related Work

Much of the recent work in human-robot collaboration focuses on the common goal of making robots a more acceptable, helpful, trustworthy, and functional part of our day-to-day life. Throughout the established literature on human-robot collaboration, a majority of the attention has been placed on providing capabilities to enable robots to adapt to their human collaborators, as opposed to providing them with the tools needed to improve their human collaborators' behaviors for more productive joint task execution.

One important trend in human-robot collaboration has been to improve robots' awareness of human behavior [2], [9], [14]. These approaches primarily focus on enabling a robot to successfully adapt and perform tasks in the presence of humans rather than enabling them to collaborate on equal footing with people. An effective approach to collaboration has been to enable the robot estimate a human collaborator's belief [15] in order to plan 'in their shoes', allowing for a better understanding of their decision-making process and the factors influencing their choices. Recent work [26] has used Inverse Reinforcement Learning (IRL) [23] to infer human behavior given a known goal. This work assumes the human holds an imperfect dynamics model for the domain, and creates a shared control scheme to invisibly correct the disparity. As our approach attributes suboptimal behavior to a human's imperfect reward model, we find applicability to scenarios (such as cognitive tasks) where shared control isn't a viable solution. Unfortunately, existing approaches do not provide mechanisms where this perspective-taking can be used to improve a human's performance and awareness on a task — rather, they mainly focus on mechanisms for allowing a robot to adapt to a human. Work by Imai and Kaneko has provided a method to estimate a human's false beliefs about a domain [19], with the intent to allow a robot to dispel said beliefs. Work by Faulkner et al. models human belief to generate minimal communication [12], enabling a robot to effectively ask for help from a human oracle, but does not investigate the reverse scenario of providing succinct help to a human agent. Implicit communication [11], [20] has also been investigated, utilizing a robot's actions to provide actionable signal about its intent in collaborative scenarios.

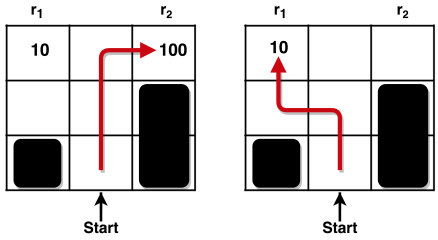
One popular approach is to develop a "theory of mind" about one's collaborator [10], [14], [28], [34] to effectively understand their knowledge state, goals, and beliefs. Work by Devin and Alami [10] estimates the information the human might be missing to minimize the conveyance of unnecessary information. In work by Leyzberg et al. [22], it is shown that personalized interactions lead to better results, while in [25] trust is better preserved and maintained by performing actions that respect a human's preferences.

During collaboration, interruptions are necessary for effective resynchronization of expectations. A great deal of work has been performed to study how [27] and when [4], [6], [31] an autonomous agent should interrupt a teammate, how to personalize interruptions [8], and even how interruptions can cause more errors in skill-based tasks [21]. **Our work addresses a crucial technical gap as it not only estimates a collaborator's belief about the reward function of their current task, but also infers the root cause for inaccuracies encoded in said belief.** Doing so provides the infrastructure needed for achieving the autonomous repair of a collaborator's policy through explanations generated online during task execution intended to illustrate and eliminate their root cause.

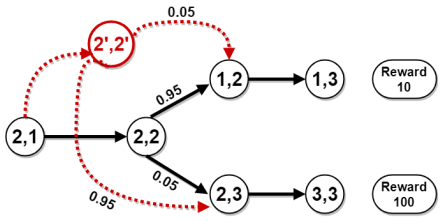
III. A Framework for Reward Augmentation and Repair through Explanation

In this section we detail the theoretical framework of *Reward Augmentation and Repair through Explanation* (RARE), wherein we utilize a Partially Observable Markov Decision Process (POMDP) coupled with a family of Hidden Markov Models (HMMs) to infer and correct a collaborator's task understanding during joint task execution. The central insight underpinning the proposed method is that sub-optimal behaviors can be characterized as an incomplete or incorrect belief about the reward function that specifies the task being performed. By proposing potential (erroneous) reward functions and evaluating the behavior of a virtual agent optimizing its policy using these functions, our approach allows a robot to determine potential sources of misunderstanding. Once a plausible reward function is discovered that explains the collaborator's behavior, a repairing explanation can be generated and provided if the benefit of correction outweighs the consequences of ignoring it.

The framework can be characterized through three interconnected components responsible for: 1) estimating a collaborator's comprehension of a domain's reward function; 2) determining a policy for trading-off between collaborative



(a) Task executions given two different comprehensions of a gridworld domain's reward function.



(b) Incorporation of reward comprehension to infer an agent's likely policy. State (2,2) indicates partial knowledge of the domain reward function, while (2',2') indicates full knowledge.

Fig. 2: An example of two possible comprehensions of a domain's reward function. (a)-left: The agent knows the true reward of the domain. (a)-right: The agent does not know about the +100 reward, and behaves rationally given this malformed reward function. (b): Latent reward comprehension variables differentiating state (2,2) and (2',2') provide a hypothesis to better explain the agent's behavior, distinguishing between the scenarios represented in (a)-left and (a)-right.

task execution and intervention; and 3) formulating corrective explanations for reward function repair. For the remainder of the section, we focus on the use case where the collaborating agent is a human and the agent employing RARE is a robot jointly executing a task with them.

A. Estimation of Reward Comprehension

The core insight of RARE is that sub-optimal behavior is an indicator of a malformed reward function being used by an otherwise rational actor. Thus, if it is possible to determine which reward function the actor is using, it will be possible to identify problematic misconceptions that may contribute to adverse behavior. As a result of this formulation, RARE necessarily assumes that the agent implementing it has a complete specification of the domain's true reward function.

To determine which components of the reward function the human collaborator is using, RARE utilizes an HMM that incorporates both state features of the world ("world features") and latent state features that indicate knowledge of corresponding components of the domain's reward function ("comprehension features"). In the example shown in Figure 2, the reward function has two components: a +10 reward for entering the top left cell and a +100 reward for entering the top right cell. The transition probabilities of a given state are directly computed from a policy trained on the reward function specified by the values of the comprehension features in the state.

We define an augmented HMM (RARE-HMM) as the 7-tuple $\lambda = \{S, O, M, \pi, A, B, \tau\}$ that estimates the likelihood of a state-action trajectory of an observed agent given a particular reward function, where:

- $S = s_0, s_1, ..., s_N$ is the finite set of states the observed agent can be in.
- $O = o_0, o_1, ..., o_M$ is the finite set of possible observations, which correspond to the effects of the action most immediately taken by the observed agent.
- M is a Markov Decision Process (S, A, T, R) where S is the set of states in the MDP, A is the set of actions an agent may take ($A = O$), T is a stochastic transition function describing the action-based state transition dynamics of the model, and R is a reward function. Intuitively, M serves as a simulator for an agent in the task domain.
- π is a policy trained to maximize reward in M.
- A is a stochastic transition matrix, indicating the transition probability from state i to state j: $A_{i,j} = P(q_t = s_j | q_{t-1} = s_i)$, where $0 \leq i, j \leq N-1$, q_t is the state at time t, and $\forall i \in [0, N], \sum_{j=0}^{N} A_{i,j} = 1$. These probabilities are drawn directly from the composition of the transition dynamics function M_T and π. In other words, A represents the transition likelihoods for an agent following policy π in M.
- B is the stochastic observation emission matrix, indicating the probability of getting observation j at time t in state i: $B_{i,j} = P(v_t = o_j | q_t = s_i)$, where $0 \leq i \leq N$, $0 \leq j \leq M$, and v_t is the observation emitted at time t. $\forall i \in [0, N], \sum_{j=0}^{M} B_{i,j} = 1$.
- τ is the distribution describing the probability of starting in a particular state $s \in S$ such that $\sum_{i=0}^{N} \tau_i = 1$.

Specifically, RARE utilizes a set of such HMMs Λ, where each member $\lambda \in \Lambda$ uses a unique reward function.

B. Collaborative Task Execution and Reward Repair

For a given collaborative task, we define the RARE agent's behaviors with a policy that solves an augmented POMDP (RARE-POMDP) defined by the 6 tuple: $(S, A, T, R, \Omega, \mathcal{O})$ where:

- S is the set of world states, consisting of both traditional features W ("world features") and additional latent features C indicating the collaborator's understanding of the domain's reward function ("comprehension features"). We formulate the set of comprehension features as a vector of boolean variables indicating whether a particular component of the reward function is known by the collaborator.

$$S = \begin{bmatrix} W \\ - \\ C \end{bmatrix}, W = \begin{bmatrix} x \\ y \\ \vdots \end{bmatrix} C = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$$

- A is the set of actions, consisting of both task-specific physical actions and reward repair-specific social actions.

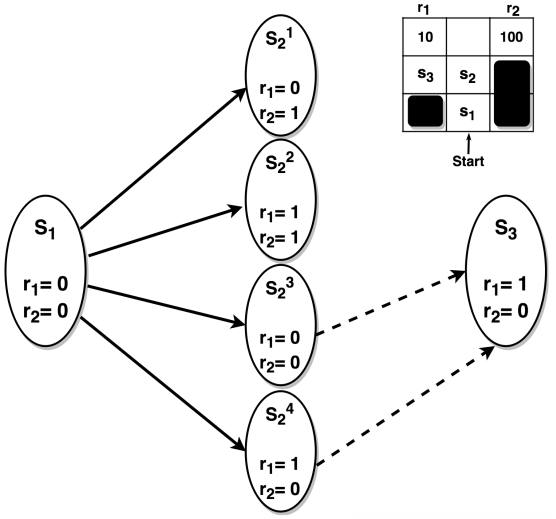


Fig. 3: Partial visualization of comprehension features for a gridworld domain with two reward factors, one at each terminal reward state. Four variants of s_2 are shown, each indicating a different level of reward function awareness. Observing an agent transition from state s_2 to s_3 provides evidence suggesting they may not know about the larger reward r_2 in the top-right, but do know about reward r_1.

- T is a transition function specifying state transitions as a function of action performed. As RARE models a collaborative process, the dynamics introduced by the collaborator's actions are also represented within this function, but are assumed to be known given known comprehension features (i.e., if the agent's reward and policy are assumed to be known, its behavior in a given state is also known).
- R is a reward function specifying the value of executing an action in a given state.
- Ω is the set of all possible observations. In a RARE-POMDP, each observation corresponds to a particular RARE-HMM being the most likely explanation for a collaborator's behavior, signaling the current state of their reward comprehension (i.e., their understanding of the reward function).
- $\mathcal{O}$ is a function describing observation emission probabilities for a given state. In RARE, the emission function must be designed to encourage congruence between a state's comprehension features and the RARE-HMM with the corresponding reward function in Ω. In other words, a RARE-HMM has higher likelihood if its reward function contains the components indicated by the current state's comprehension features.

The observation emission function presents an important design decision for implementing a RARE-POMDP in a given domain. This function provides a link between the RARE-HMMs, each representing an agent's expected behavior given a particular understanding of a reward function, and the RARE-POMDP that is being solved to maximize the success of the collaboration. In this work, we propose a softmax scoring function based on the likelihood of the collaborator's action sequence for each potential RARE-HMM. For a given observed collaborator trajectory T, RARE-HMM/observation $o_i \in \Omega$ and state $s \in S$, we propose $\mathcal{O}$ such that:

$$P(o_i|s) = \frac{\exp(P(T|o_i))}{\sum_{j=0}^{|\Omega|} \exp(P(T|o_j))}$$

Intuitively, this choice of $\mathcal{O}$ enforces that the RARE-POMDP's estimate for which reward function the collaborator is following is proportional to the likelihood that their behavior was informed by a policy derived from it. In applications where there is not a 1-to-1 correspondence between available RARE-HMMs and potential reward functions (i.e., there are not 2^n RARE-HMMs defined for n reward function components), a more clever approach may be merited.

The RARE-POMDP introduces the opportunity for the agent to make the decision to execute social actions aimed at better informing a collaborator about the domain's reward function. In other words, the agent may execute a communicative action to explicitly inform a collaborator about part of the reward function, directly changing the value of a latent comprehension feature (e.g., the knowledge of r_2's existence in Figure 3). Even though such an action may not directly advance the task toward completion, it may invariably result in higher net reward, as it can improve the collaborator's policy by informing them of high reward states or harshly penalized states that may lead to task failure.

C. Explanation Generation

The RARE framework allows an agent to estimate a collaborator's reward function during joint task execution. This is a powerful piece of information, but it is far more useful in a collaborative context when paired with actions that enable one to augment a collaborator's understanding of the task. RARE uses this information to decide what and when to communicate, updating the collaborator's reward function and policy. For our application domain, we propose an algorithm (Algorithm 1) that autonomously produces statements capable of targeted manipulation of a collaborator's comprehension features based on anticipated task failures. Future work may provide similar algorithms for providing information about non-terminal state rewards or for more generally suggesting collaborator reward function updates.

Intuitively, Algorithm 1 performs a forward rollout of a policy trained on the estimated human reward function, which may contain a subset of the information (factors) of the true reward function known to the RARE agent. As in Figure 3, the collaborator may only know of r_1, so we say it is missing the reward factor r_2. Upon completing this rollout, we also run forward rollouts for policies trained on reward functions that include one more reward factor than the human's (Figure 2). This step allows the RARE agent to find the most valuable single reward update to provide the collaborator, updating their policy by changing one reward factor at a time, following an iterative interaction pattern previously validated within HRI [3]. Finally, the update is serialized using designer-specified action [32], state [17], and reward factor description functions.

Algorithm 1: Augment Terminal-State Reward Comprehension

Input: Factored Reward Function R, Set of Policies Π Trained on Power Set of R, Estimated Human Reward Function R_h, Domain MDP $M = (S, A, T)$, Current state s_c

Output: Semantic Reward Correction

$r_c \leftarrow 0$; // Cumulative reward
$s' \leftarrow \emptyset$;
// Simulate existing human policy
$\pi_h \leftarrow$policy trained on R_h;
while *s is not terminal* **do**
    // Perform forward rollout of π_h
    $s' \leftarrow M_T(s, \pi_h(s))$;
    $r_c \leftarrow r_c + R(s, \pi_h(s), s')$;
    $s \leftarrow s'$;
$s_{h,terminal} \leftarrow s$; // Terminal state of human policy
$r_h \leftarrow r_c$;
// Find best single-comprehension-change
$\Pi_1 \leftarrow \{\pi \in \Pi \mid \pi$ trained on $R_1 \in R$ s.t. R_1 contains 1 additional factor of R^* than $R_h.\}$;
$\pi_c \leftarrow \emptyset$;
$r_\pi \leftarrow r_h$;
for $\pi \in \Pi_1$ **do**
    $s \leftarrow s_c$;
    $r_c \leftarrow 0$;
    while *s is not terminal* **do**
        // Perform forward rollout of π
        $s' \leftarrow M_T(s, \pi(s))$;
        $r_c \leftarrow r_c + R(s, \pi(s), s')$;
        $s \leftarrow s'$;
    if $r_c > r_\pi$ **then** $r_c \leftarrow r_\pi, \pi_c \leftarrow \pi$;
feedback $\leftarrow$ "If you perform {describe_action(π_h)}, you will fail the task in state {describe_state($s_{h,terminal}$)} because of {describe_reward(diff(R_h, R_π))}";
return *feedback*

IV. Experimental Validation

To quantify the viability and effectiveness of RARE within a live human-robot collaboration, we conducted a user study wherein participants had to solve a complex collaborative puzzle game – a color-based variant of Sudoku – collaboratively with a Rethink Robotics Sawyer manufacturing robot. In the sections that follow, we present results characterizing participants' perception of a RARE-enabled robot that provides guidance during complex collaborations to prevent task failure. Failure prevention was attempted by the robot by means of verbal interruptions taking place between the human's selection of a color to play and the human's placement of that color. Additionally, we investigate the role that justification plays when providing advice that directly alters the collaborator's understanding of the game.

Participants were recruited into one of two treatments that determined what the robot would communicate when interrupting a human who is about to play a move that leads to failure: a failure identification-only condition ('control') where future failures are identified but not explained, and an experimental condition ('justification') where future failures are both identified and explained to the collaborator. Participants were assigned to a third, implicit baseline condition ('no interruption') when no failures were detected and the robot did not interrupt the game.

A. Hypotheses

We conducted a human-subjects study to investigate the following hypotheses regarding RARE's application within a live human-robot collaborative puzzle-solving task:

- **H1**: Participants will find the robot more helpful and useful when it explains why a failure may occur (i.e., participants in the 'justification' condition will find the robot to be more helpful than in 'no interruption' condition and control condition.
- **H2**: Participants will find the robot to be more intelligent when it gives justifications for its actions as compared to the other conditions.
- **H3**: Participants will find the robot more sociable when it provides justifications for its failure mitigation than when it doesn't.

B. Experiment Design

To evaluate our hypotheses, we conducted a between-subjects user study using a color-based collaborative Sudoku variant played on a table with a grid overlay using colored toy blocks. Study participants were assigned into one of three conditions:

- **Control**: The robot interrupts users that are about to make erroneous block placements, indicating to them that it will cause task failure.
- **Justification**: The robot interrupts users about to make erroneous block placements, indicating that it will cause task failure and explaining which game constraint will inevitably be violated.
- **No Interruption**: An implicit condition for participants that do not commit any errors and experience interruptions by the robot.

During the game, participants place blocks concurrently with the robot (i.e., without turn-taking), until the board is filled. Participants were required to place blocks successively in the grid cells most proximal to themselves, enforcing that the final row for both human and robot were adjacent (the middle of the board). As in Sudoku, certain blocks were pre-placed on the board to limit the solution space of the task.

The robot was pre-trained on all possible solutions for the game board, making it an expert on the task. Human participants were not exposed to the board before beginning the task, and as such could be considered novices trying to solve the game online — making them susceptible to errors. During gameplay, the robot is able to verbally interrupt the human player before they place a block that will make the

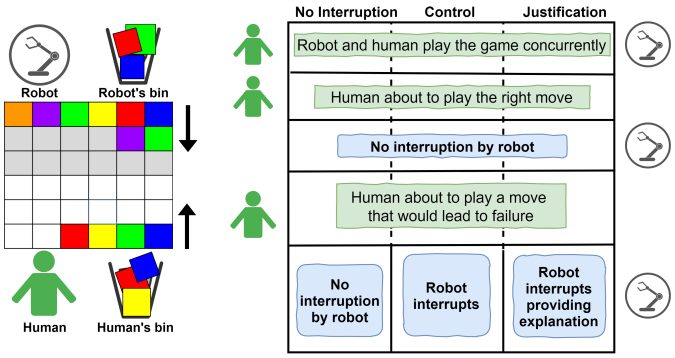


Fig. 4: (Left) Board layout for the collaborative color-based Sudoku variant. Each player concurrently fills in the three rows closest to them with colored blocks, respecting the game's constraints. The adjacency of each player's final row introduces non-trivial coordination requirements. (Right) Diagram of game flow across the three experimental conditions.

game impossible to solve, with the opportunity to provide feedback that may avoid task failure.

C. Rules of the Game

Participants must collaboratively solve a color-based 6x6 cell Sudoku variant (Figure 4), by placing colored blocks on the table until the grid is filled. There were six unique colors of block available, with a large supply of all colors available to each player. Both the participant and robot were required to place blocks from right to left, nearest-row to farthest-row, enforcing the constraint that the middle of the board is filled last (where the need for coordination is maximized). The game has two major constraints (Figure 5) limiting the gameplay decisions of both the robot and the participant.

- Row Constraint: The first constraint restricts each row of the game board to have only one of each color type.
- Adjacency Constraint: The second constraint requires that no block may have a neighbor (assuming an 8-connected grid) of the same color.

The robot and participant solve the game concurrently and independently of each other's pacing. We enforced the restriction that players must solve the row closest to them in full before moving on to successive rows, as this introduces complex coordination requirements early in the game, as early decisions will have non-obvious effects on allowable middle-row configurations. In other words, blocks placed by the robot in its third row will invariably restrict the gameplay of the participant and vice versa. Per the design of the study, the robot analyzes the gameplay decisions of the participant online and generates an interruption should they make a move that violates the constraints or inhibit successful game completion.

D. Study Protocol

Before the start of the experiment, informed consent was obtained and participants were educated about the rules of the game. We administered 1-move test puzzles, illustrating specific scenarios possible within the Sudoku variant, to verify their understanding of the game's rules and various constraints.

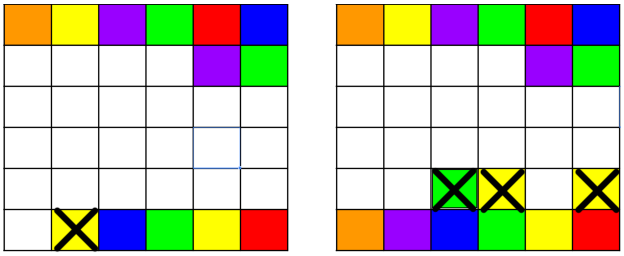


Fig. 5: Two types of violation that can occur during gameplay. Left: All colors within a row must be unique. Right: No color can be next to itself.

Both participants and the robot were both free to place blocks on the board as quickly as they were able. To play a move in the game, participants were required to: 1) Move a block from their block supply (the left-most grid of blocks in Figure 1) to a staging area (the white area directly in front of participant); 2) Solve a distractor task; and 3) Either place the staged block onto the game board or return it to the block supply and return to step 1. The staging dynamic was implemented to provide the robot with a brief moment within which it could interrupt the participant should their choice of block be determined to cause an inevitable task failure. We utilize multiplication problems as distractor tasks, though the correctness of the participant's answers was not verified.

Any blocks placed on the game board were considered final and could not be changed. If the human placed a block that prevented the game from being completed, the robot would halt the game by saying, *"I am sorry, the game cannot be solved now."* Otherwise, gameplay continued until the human and robot both solved their respective sides of the board.

At the conclusion of the game, participants were lead away from the game board to complete the a post-experiment survey and exit interview. Following the experiment, a comprehensive analysis of the dependent variables using objective measures (e.g., task completion time, idle time and number interruption) and subjective measures (e.g., Likert scale, open-ended survey questions) were used to assess the overall effectiveness of the proposed approach.

E. Implementation

Sawyer picked blocks from its supply and placed them on the board according to the game's rules. Concurrently, the robot controller implemented RARE, which monitored the current board state and human's actions, occasionally performing verbal interruptions according to the condition being run. For this game, we abstracted reward into three classes for comprehension variables: row constraint, adjacency constraint, and victory. Human-understandable feedback was generating using these with Algorithm 1. To make the game solvable quickly, we used an algorithmically predetermined board configuration to minimize the reachable states, accelerating exploration of the solution space.

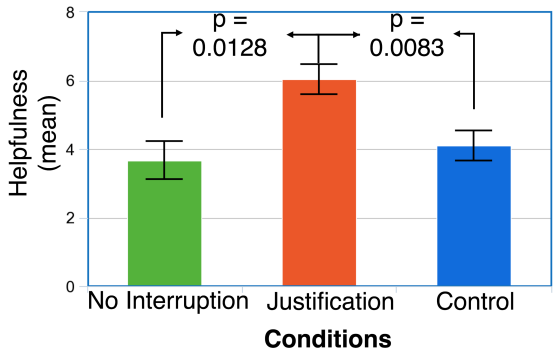


Fig. 6: Mean ratings of Helpfulness across three experimental conditions. Tukey's HSD test shows a statistically significant difference between all three conditions.

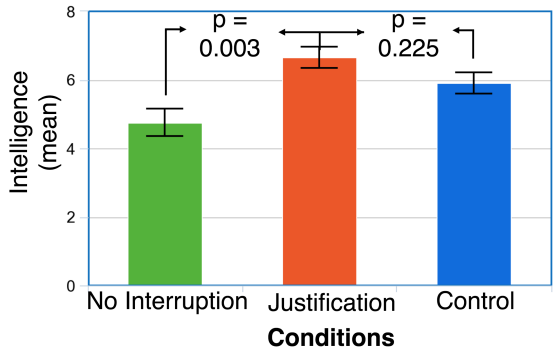


Fig. 7: Mean ratings of Intelligence across three experimental conditions. Tukey's HSD test results show a significant effect between the justification condition and the no-interruption conditions, but not between no-interruption and control.

F. Measurement

Our IRB-approved study was completed by 26 participants recruited from a university population. Participants' reported gender skewed male (65% male), and ranged in age from 18 to 30 (M = 21.87; SD = 2.93). All participants came from STEM backgrounds, and their familiarity with robots was relatively high (M=5.08, SD=1.28) on a scale from 1 to 7.

An exit-questionnaire was administered to participants after the conclusion of the game. The questionnaire was developed using questions derived from established collaborative robotics questionnaires [5], [18]. Participants were asked to rate their opinion and experience with Sawyer in 7-point Likert-scale items. Three concepts were identified which form the basis of our hypotheses, based on the previous study of shared autonomy and mixed observability of human and the agent: *Helpfulness, Sociability and Intelligence*. To determine these concepts, we first extracted the latent factors using principal component analysis (PCA). The identified factors were reduced to 11 using the Kaiser criteria, selecting factors with eigenvalues greater than 1. To spread variability more evenly across each factor, we calculate the loadings of each variable on each factor and applied varimax rotation. To identify the items that can be combined to construct a valid scale, we applied a cutoff point of correlation $r > 0.6$ to the factor matrix.

Sociability was comprised of questions measuring participants' opinions about Sawyer with respect to the interaction's naturalness, pleasantness, and comfort ($\alpha = 0.8557$).

Helpfulness was comprised of questions measuring participants' opinions about how useful and informative Sawyer was during the interaction and its ability to help ($\alpha = 0.83$).

Intelligence was comprised of questions measuring participants' opinions about how intelligent and knowledgeable Sawyer was ($\alpha = 0.8734$).

V. Results and Discussion

A. Analysis

There were no anomalies or outliers detected in our combined data set for any of the three concepts, but the datasets were positively skewed. We did not observe any multimodalities in the distribution of data. We conducted an ANOVA to test effects across our experimental conditions with respect to perceptions of *Sociability, Helpfulness, and Intelligence*.

We found a significant effect from the 'justification' condition on perceived helpfulness ($F(2, 23) = 7.23, p < 0.004$), **confirming H1**. Post-hoc comparisons using Tukey's HSD test (Figure 6) revealed that the justification condition resulted in a significantly different level of Helpfulness as compared to the control condition ($p < 0.009$) and the no-interruption condition ($p = 0.013$).

We also found a significant effect caused by the justification condition on our measure for intelligence ($F(2, 23) = 6.99, p < 0.005$), **confirming H2**. Post-hoc comparisons using Tukey's HSD test (Figure 7) revealed that the justification condition resulted in a significantly different level of our perceived intelligence measure as compared with the no-interruption condition ($p = 0.003$), but not with the control condition ($p = 0.225$). Hence, we cannot dismiss the null hypothesis that a robot notifying a collaborator of a bad action choice may not be differently perceived if it also offers justification for its advice.

No significant effects were found with respect to perceptions of sociability as a function of experimental condition ($p = 0.1$), thus **we cannot validate H3**.

Objectively, we observed that there were more terminations of the game during the control condition as compared to the justification condition ($8/10$ vs $2/10$) which we did not anticipate when designing our experiment. As the robot preempts human actions that would lead to task failure in both conditions, we anticipated that the our control condition (notification of inevitable failure without justification) might lead to longer completion times. To understand the behavior of participants who ignored the robot's warning, we looked to the open-answer questions in our exit survey.

One of the two participants that had their game terminated due to invalid block placement in the justification condition indicated that they were too involved in the game and did not listen to Sawyer's advice and warnings:

"I was so much involved in completing the game, I completely missed [the warning] from the robot — I just heard some sound from the robot and did not realize what it was saying..."

The other participant indicated that they started to think of Sawyer as a competitor and did not listen to its advice, despite being briefed on the collaborative nature of the game at the onset of the experiment:

"As soon as the game began, I forgot it was a collaborative game and I became competitive and was not sure of advice given by Sawyer"

In the control condition, the survey responses painted a clear picture for the terminations — **participants did not trust Sawyer when it indicated that the human was about to make a move that would cause the task to eventually fail, when it did so without further explanation**. They were confused why the move was not valid, even though it looked valid to them. They were skeptical with respect to Sawyer who was not providing accompanying justification for its judgment of their move, as evidenced by the following quotes from participants' survey responses:

- *"Sawyer wasn't forceful enough and was not giving me the reasons why the move was wrong. So I couldn't trust him"*
- *"Response looked like hard coded and I did not find the reason to think that Sawyer was addressing to me"*
- *"I felt that Sawyer was a robot that is good but I didn't know what his purpose was ... I feel he should have been more forceful in stopping me doing the wrong moves."*
- *"I did not believe it as it did not give details regarding the wrong step"*

We also found evidence in the post-experiment surveys supporting the notion that **providing justification alongside reward feedback leads to a more positive user experience**. Many participants found easier to trust Sawyer when it was providing an explanation alongside its advice. We also saw evidence that the behavior in the justification condition was affecting the way participants played, an important result.

- *"He was forward predicting the movement of the game and telling me why my move was not right even though it was the right move. I was able to trust him easily when he gave the reasons"*
- *"It helped me make sure that I made the correct decisions"*
- *"I learnt to think of moves ahead when Sawyer helped me once with the game."*
- *"Sawyer's input made me question my understanding of the game"*

Thus, we can conclude from the qualitative and quantitative results of our user study that RARE provides tangible subjective and objective benefits during human-robot collaboration. Our experimental results further show improvements beyond standard failure mitigation techniques. Our results highlight that justification is an important requirement for a robot's corrective explanation. Hence, we validate that our contribution is not a solution in search of a problem, but addresses an important, underexplored capability gap in the HRI and Explainable AI literature.

B. Opportunities for Future Work

Our proposed framework allows an agent to estimate and provide corrections to a collaborator's reward function during joint task execution. RARE's effectiveness stems from its ability to discover the root cause for an agent's suboptimal behavior and provide targeted, interpretable feedback to address it. One of the drawbacks of RARE is that the formulation of reward factors by way of comprehension features causes the state space to explode combinatorially, with non-trivial reward functions causing RARE to easily become intractable.

There are many potential approaches for addressing this problem of scalability: 1) Attention mechanisms and priors to reduce comprehension features (i.e., making a priori assumptions about what one's collaborator knows); 2) State abstractions to reduce state space [1]; and 3) Reward function abstractions (i.e., removing the naive independence assumption of rewards across states), approximations/simplifications, or using a subset of potential reward function candidates.

Furthermore, in our implementation the RARE framework estimates only *missing rewards* from the user's comprehension of a domain's true reward function. We are not considering the cases where the user has an imagined reward not truly present in the true reward function, or in other words, where the user erroneously includes incorrect or non-existent reward signal in their comprehension of the domain.

Based on the exit interviews of participants who ignored the robot's advice due to over-engagement in the game, where participants said they were too busy to listen, a promising direction for future work also includes investigating different modalities for conveying reward repair information (e.g., incorporating nudging theory for non-invasive corrections).

Finally, we have considered only a single RARE agent (expert) and a collaborator (novice). Natural extensions of this work include relaxing assumptions about the RARE agent's knowledge of the true reward function (e.g., can RARE be improved to enable two RARE agents with complementary reward functions learn a stronger joint reward function from each other's feedback) or extending the work to larger teams.

VI. Conclusion

In this work we proposed Reward Augmentation and Repair through Explanation, a novel framework for estimating and improving a collaborator's task comprehension and execution. By characterizing the problem of suboptimal performance as evidence of a malformed reward function, we introduce mechanisms to both detect the root cause of the suboptimal behavior and provide feedback to the agent to repair their decision-making process. We conducted a user study to investigate the effectiveness of RARE over a standard failure mitigation strategy, finding that **RARE agents produce more successful collaborations and are perceived as more helpful, trustworthy, and as a more positive overall experience.**

References

[1] D. Abel. A theory of state abstraction for reinforcement learning. In *Proceedings of the Doctoral Consortium of the AAAI Conference on Artificial Intelligence*, 2019.

[2] R. Alami, A. Clodic, V. Montreuil, E. A. Sisbot, and R. Chatila. Toward human-aware robot task planning. In *AAAI spring symposium: to boldly go where no human-robot team has gone before*, pages 39–46, 2006.

[3] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–149. ACM, 2018.

[4] S. Banerjee, A. Silva, K. Feigh, and S. Chernova. Effects of interruptibility-aware robot behavior. *arXiv preprint arXiv:1804.06383*, 2018.

[5] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[6] D. Brščić, T. Ikeda, and T. Kanda. Do you need help? a robot providing information to people who behave atypically. *IEEE Transactions on Robotics*, 33(2):500–506, 2017.

[7] J. Casper and R. R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3):367–385, 2003.

[8] Y.-S. Chiang, T.-S. Chu, C. D. Lim, T.-Y. Wu, S.-H. Tseng, and L.-C. Fu. Personalizing robot behavior for interruption in social human-robot interaction. In *Advanced Robotics and its Social Impacts (ARSO), 2014 IEEE Workshop on*, pages 44–49. IEEE, 2014.

[9] M. Cirillo, L. Karlsson, and A. Saffiotti. Human-aware task planning for mobile robots. In *Advanced Robotics, 2009. ICAR 2009. International Conference on*, pages 1–7. IEEE, 2009.

[10] S. Devin and R. Alami. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, pages 319–326. IEEE, 2016.

[11] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 51–58. ACM, 2015.

[12] T. K. Faulkner, S. Niekum, and A. L. Thomaz. Robot dialog optimization via modeling of human belief updates. 2017.

[13] J. Forlizzi. How robotic products become social products: an ethnographic study of cleaning in the home. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 129–136. ACM, 2007.

[14] O. C. Görür, B. S. Rosman, G. Hoffman, and S. Albayrak. Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. 2017.

[15] J. Guitton, M. Warnier, and R. Alami. Belief management for hri planning. *BNC@ ECAI 2012*, page 27, 2012.

[16] B. Hayes and B. Scassellati. Challenges in shared-environment human-robot collaboration. In *"Collaborative Manipulation" Workshop at the 8th ACM/IEEE International Conference on Human-Robot Interaction.*, page 8, 2013.

[17] B. Hayes and J. A. Shah. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 303–312. ACM, 2017.

[18] G. Hoffman. Evaluating fluency in human-robot collaboration. In *International conference on human-robot interaction (HRI), workshop on human robot collaboration*, volume 381, pages 1–8, 2013.

[19] J.-I. Imai and M. Kaneko. Development of robot which recognizes user's false beliefs using view estimation. In *World Automation Congress (WAC), 2010*, pages 1–6. IEEE, 2010.

[20] R. A. Knepper, C. I. Mavrogiannis, J. Proft, and C. Liang. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, pages 283–292. ACM, 2017.

[21] B. C. Lee and V. G. Duffy. The effects of task interruption on human performance: a study of the systematic classification of human behavior and interruption frequency. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 25(2):137–152, 2015.

[22] D. Leyzberg, S. Spaulding, and B. Scassellati. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 423–430. ACM, 2014.

[23] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

[24] S. Nikolaidis and J. Shah. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 33–40. IEEE Press, 2013.

[25] S. Nikolaidis, Y. X. Zhu, D. Hsu, and S. Srinivasa. Human-robot mutual adaptation in shared autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 294–302. ACM, 2017.

[26] S. Reddy, A. Dragan, and S. Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1461–1472. Curran Associates, Inc., 2018.

[27] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans?: strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 109–116. ACM, 2009.

[28] B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[29] M. Spenko, H. Yu, and S. Dubowsky. Robotic personal aids for mobility and monitoring for the elderly. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(3):344–351, 2006.

[30] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy. Asking for help using inverse semantics. 2014.

[31] J. G. Trafton, A. Jacobs, and A. M. Harrison. Building and verifying a predictive model of interruption resumption. *Proceedings of the IEEE*, 100(3):648–659, 2012.

[32] N. Wang, D. V. Pynadath, and S. G. Hill. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 997–1005. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[33] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada, K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba. Home-assistant robot for an aging society. *Proceedings of the IEEE*, 100(8):2429–2441, 2012.

[34] Y. Zhao, S. Holtzen, T. Gao, and S.-C. Zhu. Represent and infer human theory of mind for human-robot interaction. In *2015 AAAI fall symposium series*, volume 2, 2015.