Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming

Matthew B. Luebbers*, Aaquib Tabrez*, Kyler Ruvane*, and Bradley Hayes
University of Colorado Boulder
{matthew.luebbers, mohd.tabrez, kyler.ruvane, bradley.hayes}@colorado.edu

***Abstract*—Justification is an important facet of policy explanation, a process for describing the behavior of an autonomous system. In human-robot collaboration, an autonomous agent can attempt to justify distinctly important decisions by offering explanations as to why those decisions are right or reasonable, leveraging a snapshot of its internal reasoning to do so. Without sufficient insight into a robot's decision-making process, it becomes challenging for users to trust or comply with those important decisions, especially when they are viewed as confusing or contrary to the user's expectations (e.g., when decisions change as new information is introduced to the agent's decision-making process). In this work we characterize the benefits of justification within the context of decision-support during human-robot teaming (i.e., agents giving recommendations to human teammates). We introduce a formal framework using value of information theory to strategically time justifications during periods of misaligned expectations for greater effect. We also characterize four different types of counterfactual justification derived from established explainable AI literature and evaluate them against each other in a human-subjects study involving a collaborative, partially observable search task. Based on our findings, we present takeaways on the effective use of different types of justifications in human-robot teaming scenarios, to improve user compliance and decision-making by strategically influencing human teammate thinking patterns. Finally, we present an augmented reality system incorporating these findings into a real-world decision-support system for human-robot teaming.**

I. Introduction and Motivation

Many works in the explainable AI (xAI) literature have illustrated the benefits of illuminating the black box of AI decision-making for end users interacting with autonomous and robotic agents [63, 25, 5]. Various xAI techniques facilitate better transparency into collaborative robots' choices, improving trust, interpretability, and user acceptance [8, 56, 17, 42]. However, if explanations are given at inopportune times with poor context, they can produce the opposite effect [30]. Furthermore, different explanation content can have differing effects on a human collaborator's mental model, which can impact their behavior [7, 40]. In this work, we hypothesize that since human collaborators have limited cognitive bandwidth to process explanations, it is best to time them strategically for maximum impact on improving understanding and behavior. We also propose that the content and manner in which the explanations are given should be tailored to a collaborative context to encourage the desired effect on a human teammate.

* These authors contributed equally to this work

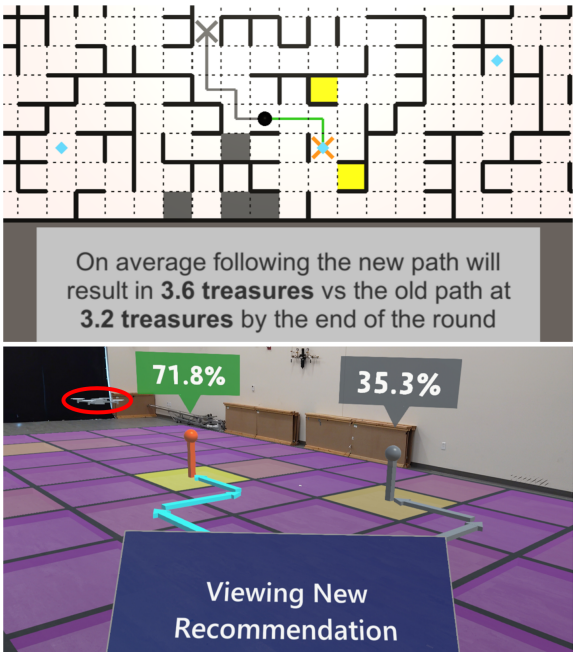

Fig. 1: *Top:* a counterfactual policy-based justification provided by drones (blue diamonds) to the human in a collaborative 2D treasure hunting game. *Bottom:* a counterfactual environment-based justification showing the relative percentages of finding a target, provided by a drone (circled in red) in an augmented reality navigation interface. Both justifications are attempting to explain to a user why they should take a new (colored) recommended path, rather than the old (gray) path.

In collaborative human-robot interaction tasks, accounting for a human in a multi-agent planner is challenging due to the innate unpredictability and opacity of the human's decision-making [51, 27]. Therefore, having a robotic teammate also act as a decision support system for the human, suggesting actions for the human to perform while itself working towards a shared task, is helpful for alleviating this unpredictability [59, 57, 8, 54]. With this type of interaction, it is crucial that autonomous agents **justify** their behavior or suggestions when they deviate substantially from the human teammate's expectations.

We define justifications in this context as explanations timed appropriately to instances of expectation mismatch, with the intent of convincing or influencing a human agent. For example, in a human-robot collaboration scenario where a robotic agent is providing navigation recommendations, a

sudden change in the recommended direction may appear confusing and strange to the human teammate, and is likely to be disregarded [59]. A justification (see Fig. 1 for examples) provided in this context serves to convince the human teammate of the utility of the previously difficult to interpret recommendation. Our work addresses two research questions: 1) When are such justifications most impactful and useful? And 2) What information should be presented in justifications to improve human teammate decision-making and behavior?

The core contributions of this work are as follows:

- A novel mathematical framework, informed by value of information theory, to decide *when* a robot collaborator should justify its recommendation to a human teammate, validated by an expert-feedback case study for determining the utility of justification timing strategies.
- A methodological characterization of four different types of justification, derived from established features in xAI literature, along with a validation and analysis of these justification types via an online human subjects study.
- A set of actionable design recommendations and implementation strategies for the use of justifications in human-robot interaction, taking into account differing levels of human and robot decision-making competence, along with an augmented reality interface showcasing these design principles for practical applications.

II. Background & Related Work

Explainable AI and Human-Robot Interaction: Recent research on shared mental models within human-robot collaboration has shown the importance of explainability for enhancing interaction efficiency, fluency, and safety [50, 58, 63]. This is particularly relevant in the context of model reconciliation, where mismatches in expectations can lead to catastrophic failures [8, 7]. Explainable AI can help bridge the gap between human and robotic agents by making complex models more understandable, allowing for faster debugging and failure recovery, ultimately improving joint performance [28, 45, 63].

As such, it is important for robotic agents to be able to effectively communicate and explain their decision-making rationale to human collaborators, with awareness of how these explanations influence and affect team dynamics. Moreover, research has also shown that people trust autonomous agents more when they convey their decision-making process [66, 37]. Robots with this explanation-providing capability are generally perceived to be more helpful and transparent [57]. Conlon et al. [10] show that when a robot provides a self-assessing explanation, operator trust more appropriately aligns with robot ability, leading to increased performance and trust.

Explanation Strategies: Research in two areas of explainable AI are particularly relevant to explanation generation: methods that explain how a learned model functions (explainable ML) and methods that produce explainable agent behavior during human-in-the-loop interaction [55]. Explainable ML methods are often aimed at helping developers interpret complex classifiers by illustrating how individual parameters impact model output. Popular techniques include local approximations like SHAP [41], model-agnostic methods like LIME [45], and visualizations like Grad-CAM [49].

Explainable behavior methods attempt to make the intentions of robotic agents clearer to humans by improving metrics like explicability [38], predictability [6], or legibility [18]. Research has demonstrated that people dislike inexplicable behavior from robots, rating it as frustrating, and leading to mistrust of the robot [63, 2]. Robot behavior that attempts to align itself with human expectations often must sacrifice optimality to achieve high explicability. In Tabrez et al. [59], participants in a collaborative search scenario expressed a preference for explanations from an autonomous agent when its behavior was unexpected or confusing. These explanations, provided they are contextualized properly to mismatches in human and robot expectation, can serve as a bridge between explicability and optimality: alleviating the negative effects of inexplicable but optimal robot behavior, and building trust in the system over time.

Explanations as Justification: This work focuses on the strategic use of explanations as justification in human-robot teaming. This involves timing explanations to an instance of expectation mismatch between humans and robotic agents, with the goal of influencing a human teammate. Correia et al. [11] found that using justification as a recovery strategy for robot failures can mitigate the negative perception of those failures. Prior work has focused on using justification to explain why a decision is good or bad, without necessarily aiming to give an explanation of the decision-making process [19, 57]. In this work, we introduce and analyze different types of justifications aimed at addressing both of those goals.

III. Definition of Application Domain

To ground and evaluate our contributions, we utilize a multi-target search and retrieval problem as a representative human-robot teaming application. This multi-goal, multi-agent planning domain includes agents with heterogeneous capabilities operating under partial observability.

We utilize an experimental paradigm previously established by Tabrez et al. [59], which assumes two distinct classes of heterogeneous agents working toward a multi-objective task (e.g., search and recovery): autonomous agents (information-gathering agents that move through the environment and take sensor observations) and human agents (interactive agents that can directly affect the environment state with their actions and complete objectives, such as collecting a sample) in a partially observable domain. In this paradigm, humans serve as interactive agents that receive action recommendations from autonomous information-gathering agents that typically have access to features the interactive agents cannot directly perceive. The decision-making process for each class of agent is codified by a separate Markov Decision Process (MDP):

- Autonomous agent MDP, M_r, is defined by the 4-tuple: (S_r, A_r, T_r, R_r), where S_r is the set of states in the MDP, A_r is the set available actions, T_r is a stochastic transition function describing the model's action-based

state transition dynamics, and R_r is the reward function $R_r : S_r \times A_r \times S_r \rightarrow \mathbb{R}$.

- Recommendations for human agents are generated using an MDP model of the human M_h defined by a 4-tuple (S_h, A_h, T_h, R_h).

Environmental uncertainty over task-relevant variables (e.g., whether a location contains a buried sample) is characterized by a dynamically-updating probability mass function (PMF). This PMF serves as a shared utility function common to all agents (both human and autonomous), and can be communicated to human teammates as it changes in response to autonomous agent observations to provide insight into the agent's policy (additional detail provided in Section V-A).This relationship can be seen in Fig. 2.

In the multi-target search task, the PMF is in essence a heatmap representing the probability at each location for finding a target. The autonomous agent MDP M_r generates optimal moves for these information-gathering agents to attempt to collapse the uncertainty of that PMF by locating targets via sensor observations. Meanwhile, the human MDP M_h generates recommendations for the human agent to follow to achieve the task goals, constantly updating based on the most recent PMF.

The novel justification framework evaluated by our experiment was situated within the context of a human-drone collaborative search task, an established evaluation domain for decision support [59]. Fig. 2 shows the interaction flow of the task. In this section, we will use the circled letters in the diagram to walk through its implementation.

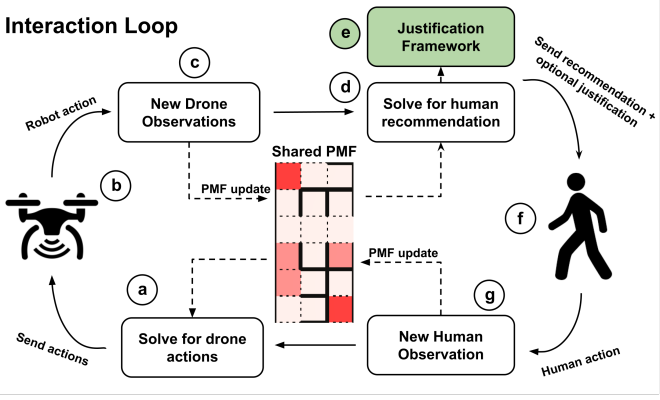

Fig. 2: The loop describing the human-drone interaction with shared PMF in our domain. The Justification Framework, the primary contribution of this work, is highlighted in green.

To start, drones solve for their next actions (a) using the MDP M_r; in our domain each drone is assigned its own segment of the environment to cover to ensure uniform search coverage. As the drones take their actions (b), they observe noisy sensor readings over the cells they fly over to attempt to detect targets (c). Using these readings, the shared PMF undergoes a Bayesian update. Next, the system calculates a recommendation for the human using M_h (d). The system determines whether a justification is needed, and if so, generates one (e); the justification framework (the primary contribution of this work) is described in detail in Sections IV and V. The human's next recommendation and optional justification are sent to the human, who then takes their next action (f). Based on the system's observation of the human action, the PMF and state is updated again (g), and the cycle returns to (a).

IV. Justification Framework: Timing

In this section, we address the question of "when" justification should be provided within human-robot teaming scenarios, and present a novel framework for the timing of justifications based on value of information theory. Throughout this section, we focus on the use case where the collaborating agent is acting as a decision support system, providing recommendations to a human agent who can either comply with or reject them.

A. Spectrum of Justification Timing Strategies

Prior work has shown that in collaborative human-robot interaction, humans are highly influenced by the timing and frequency of those interactions [30]. To examine the question of when and how frequently justifications should be presented, we start by anchoring the range of possible actions at the two extremes: never justifying or always justifying.

There are two general criteria that would render a justification unnecessary within a human-robot collaboration. 1) there are no actionable consequences stemming from the recommendation to be justified, or 2) the robot's recommendations are generally accepted and trusted without scrutiny [16]. In most adaptive autonomy use cases, the second criterion is rarely met, especially in high uncertainty environments [46, 59]. Prior research has found that whenever there is a misalignment of expectations between human and autonomous teammates, explanations are expected to be provided [59, 7]. These expectation mismatches can stem from a variety of causes, including sudden changes in recommendation or a recommendation based on environment data that is unknown to the human [8]. Trust and reliance in these systems deteriorate when they lack the capability to justify their recommendations in the presence of such mismatches [11]. In these scenarios, never justifying is undesirable.

On the other hand, always justifying is ill-suited for human-agent collaboration. Prior research has shown that administering too many queries increases frustration and irritation in users [4]. Justifying too frequently can lead to habituation, as repeated explanations reduce user responsiveness to them [61, 23, 24, 33]. Thus, always justifying is also undesirable.

B. Strategically Timing Justifications: Value of Information

Even though justifications have benefits, agents should provide them strategically to take advantage of them efficiently. As there is a direct cost of increased workload and habituation inherent to providing an explanation to users, justification should only be made when the value exceeds the cost. We utilize value of information (VOI) theory [32] to decide how much value a specific justification may add.

Value of Information. VOI is typically used in autonomous systems contexts to maximize the information that a system can gather or observe by using a "pull" communication pattern, where a requesting agent (usually an autonomous system) formally weighs the cost to query a responding agent (usually a human) to provide additional information [36].

However, as we are operating within the context of conveying an explanation to a human agent autonomously, we adopt VOI in a "push" communication pattern, where an information-providing agent (robot teammate) formally weighs the cost to a receiving agent (a human) in parsing that information, along with the cognitive burden of interrupting their current task [4].

Justification Framework. Using the human MDP M_h described in Section III, our framework constructs an optimal policy for the human π_h^*. However, this optimal recommended policy is not necessarily agreed upon by the human and the autonomous agents since they may have differing reward functions. Therefore it is necessary for the system to model the human and estimate what their π_h should be.

- $\widehat{\pi_h^*}$ is a human's optimal policy as derived from the human's own internal reward function $\widehat{R_h}$ and operating using their world model $\widehat{M_h}$. The notation '$\widehat{--}$' denotes that the variable in question is derived from the human's internal model of the world, which is latent to the system and must be estimated.
- π_h^* is the system's optimal policy for the human derived from R_h, the system's model of the human's reward function and its model of the human MDP M_h. The policy recommendation can change based on receiving new information (e.g., new sensor readings).

When there is perfect synergy between the human and the system (a shared mental model), these two policies will be the same ($\widehat{\pi_h^*} = \pi_h^*$). However, the human's and the system's understanding of the optimal policy will drift as the system receives new information and makes updates to π_h^* while the human makes potentially different choices while using out-of-date information, leading to a mismatch in the mental model.

The human and the autonomous agent will have two separate understandings of the expected reward for following a given policy starting from a state s:

- $\mathbb{E}_{\pi_h^*,s}(R_h)$ is the expected reward the *system* expects the human to receive by following the recommended policy.
- $\mathbb{E}_{\widehat{\pi_h^*},s}(\widehat{R_h})$ is the expected reward the *human* expects to receive by following their own policy.

Justification is needed when the autonomous agent's recommendation appears unintuitive or confusing to a user. We hypothesize that the two primary reasons for this confusion are 1) an explicit mismatch in the expected reward, or 2) a mismatch in the sequence of states that are expected to be visited even in the case of identical expected reward.

The first contributor is the mismatch in expected reward and is formalized as:

$$\mathcal{D} = |\mathbb{E}_{\pi_h^*,s}(R_h) - \mathbb{E}_{\widehat{\pi_h^*},s}(\widehat{R_h})| \tag{1}$$

Where $\mathcal{D}$ is a scalar representing the difference in the robot's expected reward and the human's expected reward from following their respective policies for the human agent. To formalize the second contributor, it is useful to define two possible trajectories for the human.

- ψ_h denotes the sequence of states the *system* thinks the human should traverse, obtained from a rollout of π_h^* starting from current state s.
- $\widehat{\psi_h}$ denotes the sequence of states the *human* thinks the human should traverse, obtained from a rollout of $\widehat{\pi_h^*}$ starting from current state s.

The expected mismatch in path is defined as a distance function between the two paths:

$$\mathcal{T} = dist(\widehat{\psi_h} - \psi_h) \tag{2}$$

Here, $\mathcal{T}$ is a scalar representative of the difference between the robot's recommended path and the human's expected path. We define the value of a justification, V($\mathcal{J}$), as a piecewise linear filter with three components:

$$V(\mathcal{J}) = max \begin{cases} \alpha * \mathcal{D} \\ \beta * \mathcal{T} \\ \gamma * \mathcal{D} + \kappa * \mathcal{T} \end{cases} \tag{3}$$

α, β, γ, and κ are tunable hyper-parameters. The first component of Eq. 3 captures the mismatch in the expected reward, the second captures the mismatch in the expected path, and the third provides a more comprehensive filtering criteria based on a linear combination of the two. The three filters combine to create an expressive notion of the value of a potential justification.

This justification to a user comes at a cost $C(\mathcal{J})$, which is highly dependent on the particular task and mode of communication, and should tuned separately per domain. A justification should only be triggered if the expected benefit to the user is higher than the justification cost.

$$V(\mathcal{J}) - C(\mathcal{J}) > 0 \tag{4}$$

In human-robot teaming scenarios, as the mismatch between the robot's recommendation and human mental model increases, the usefulness of the robot's recommendations decrease. VOI can be used to determine the trade-off between providing justification to bridge the gap and making the recommendations more useful.

Additional Implementation Details. Here, we present additional details about how we applied this framework to our domain. The value of a potential justification relies on the human's internal policy $\widehat{\pi_h^*}$ and the system's recommended policy for the human π_h^*. Since the human's internal policy is latent from the perspective of the system, we infer the human's most likely reward function $\widehat{R_h}$ based on the information they can observe, and derive their policy $\widehat{\pi_h^*}$ assuming that humans optimize expected reward given their current reward knowledge: a common practice within inverse reinforcement learning and preference learning literature [57, 47]. Since the

only reward information humans receive is communicated via the robots, we update the human's reward function $\widehat{R_h}$ and resultant policy $\widehat{\pi_h^*}$ whenever the robot provides a communicative update, using the reward content of that update as an approximation of the human reward knowledge (i.e., using π_h^* from the last recommendation received by the human, at a previous timestep). The human's desired path $\widehat{\psi_h}$ is estimated using π_h^* from that previous timestep.

The specific implementation for our domain of the distance function in Eq. 2 to find $\mathcal{T}$ uses an XOR of states in the human's expected path $\widehat{\psi_h}$ and the states in the new recommended path ψ_h. Simply put, the difference function takes into account states that are visited by one of the compared trajectories, but not both. Prior research has shown that people are more concerned by actions that are nearer to them [59, 43]. With that in mind, we weight differences higher the closer they are to the human's current location.

$$\mathcal{T} = \sum_{s' \in \psi_h \oplus \widehat{\psi_h}} \gamma^{d(s', s_h)} \tag{5}$$

The distance function is the sum of a tuned discount factor γ raised to the Euclidean distance $d(s', s_h)$ between a state s' and the human's current state s_h ($d(s', s_h)$) for all states s' in the XOR set $\psi_h \oplus \widehat{\psi_h}$.

We combine the scalar state difference $\mathcal{T}$ with the scalar reward difference $\mathcal{D}$, as described in Eq. 1, and tune the relevant hyperparameters in Eq. 3 to create an appropriate function for the value of justification $V(\mathcal{J})$, justifying whenever it exceeds the cost $C(\mathcal{J})$, tuned for our domain.

C. Justification Timing Case Study

We validate our VOI-based timing mechanism for offering justifications through a within-subjects expert-feedback case study (n=10) where participants (graduate students in the fields of robotics and human-computer interaction) watched video of three playthroughs of a treasure hunt game (shown in Fig. 1-top) with differing justification timing strategies. In this partially observable maze-like domain, players must uncover as many hidden treasures as possible in a limited number of turns, aided by autonomous drone teammates who explore the maze and provide continually updating recommendations based on their noisy 'treasure detector' sensor readings.

The video paused periodically during trials at moments where a justification (Fig. 1-top) could be offered. The experts were asked at each pause how useful the addition of a justification at that point in the game would be, on a scale from 1 (not useful at all) to 5 (very useful), similar to Cruz et al. [12].

Each 21-turn long playthrough utilized one of three timing strategies, presented in a random order: justifying once every turn (21 justifications), justifying at regular intervals of once every four turns (5 justifications), or justifying based on the proposed VOI-based mechanism (5 justifications). We hypothesized that users would find strategically timed VOI justifications to be more useful than constant or timed-interval justifications.

As shown in Table I, we found that strategic justification led to the highest average perceived usefulness rating, showing that it is not only preferable to justify less frequently, but also that the specific timing of justifications to periods of high mismatch in expectations is preferable to a similarly infrequent justification strategy.

	Always	**Interval**	**VOI-strategic**
Usefulness Mean	2.34	2.74	4.16
Usefulness SD	1.47	1.31	0.74

TABLE I: Means and standard deviations of rated usefulness of justification timing (on scale of 1-5) per timing strategy.

V. Justification Framework: Content

In this section, we investigate the *content* of effective justification. Drawing from previous works in explainable AI [48, 45, 1], we introduce four broad categorizations of justifications using a 2x2 cross of *environment-centric* vs. *policy-centric* and *local* vs. *global*.

The first axis of the 2x2 cross, *environment-centric* vs. *policy-centric*, determines whether the justification is grounded in features from the environment that influence the policy, or features of the resultant policy itself. As an example, an algorithm recommending a location for a new wind turbine might provide the average wind speed at various prospective locations as an *environment-centric* justification for those locations. Alternatively, it could provide the expected power produced in a year if a recommended location was chosen, contrasted with the expected power produced if alternative locations were chosen as a *policy-centric* justification.

The second axis, *local* vs. *global*, determines whether the explanation is grounded in a localized, short-horizon context, or a global, long-horizon context. While a *local* justification may focus on the sub-goals and immediate rewards of a given task, a *global* justification would give a broader overview of the end goal of a domain.

All justifications in our framework are structured counterfactually, comparing the recommendation expected by the human, derived from a model of their own policy $\widehat{\pi_h^*}$, to the current recommendation actually given to the human by the robot derived from π_h^*. Counterfactual explanations are broadly defined as answers to contrastive questions of the form "Why did outcome P happen rather than outcome Q? [62]" These explanations can be conveyed via natural language or visually. Counterfactuals have shown usefulness for model debugging and failure recovery, as these types of explanations provide contextual information about a model's internal reasoning [22, 9, 60].

The following four proposed types of features used in a justification vary along a spectrum of interpretability and comprehension for its users [15].

C1. Environmental Features: These types of features provide a sense of interpretability for users, as they get quick insight into the robot's decision-making rationale.

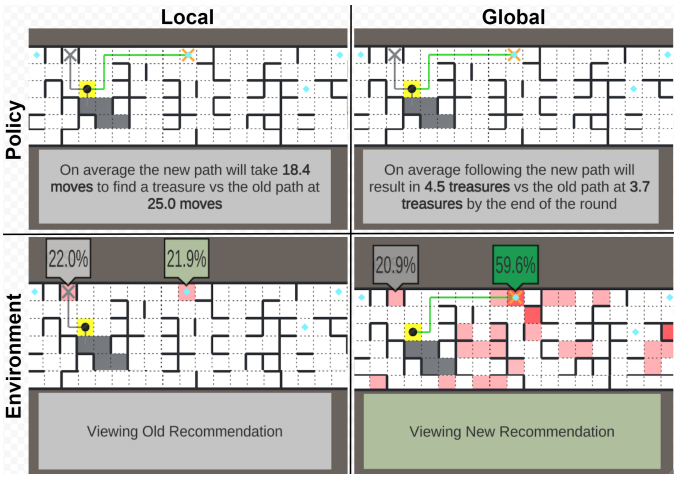

Fig. 3: The four types of characterized justifications, given during the same gameplay scenario in the treasure hunt domain. Note that the percentages shown on the map in both environment-based justifications involve alternating visually between the old and new probabilities every 1.5 seconds. For simplicity, only the old probabilities are shown for 'environment local' and the new probabilities for 'environment global' in this figure.

C2. Policy Features: These features lack in interpretability, since they don't provide any insight into the robot's rationale, but they are highly comprehensible, as the user can easily compare the end results of the agent's decision-making.

C3. Local Features: Humans are bounded by a limited cognitive capacity [53], and tend to prioritize short-term rewards in their own reasoning (e.g., Stanford marshmallow experiment [43]). Therefore, local features provide a mix of short-sighted interpretability and compliance characteristics.

C4. Global Features: Global features sacrifice precision for high comprehensibility, succinctly conveying the robot's long-term policy with human-understandable explanations tied to the success criteria of the task itself.

A. Framing Justifications for Search Tasks

We frame the four proposed justification types, built from the 2x2 cross, in the context of a multi-target search task which utilizes a dynamically updating probability mass function (PMF) as the primary element of the feature space, a common practice in search and rescue operations [20, 64, 65]. The PMF is a discrete mapping of locations to the probability of a target being found at the location. It is, in essence, a heatmap representing the likely locations of targets across the environment. As information is gathered through environmental exploration, the PMF is updated via Bayes' Rule.

To estimate mental model divergence over time, the system estimates the human's policy $\widehat{\pi_h^*}$ by using the last recommendation given to the human by the robot π_h^*, taken from a previous timestep. This leverages the assumption that the human teammate's mental model is aligned with the most recent guidance they have received from the system, with divergence occurring in the interval between justifications. To repair this divergence, four types of justification can be used:

Environment-centric Global. This justification is conveyed visually by converting the current PMF to a heatmap, with a color gradient from white to red representing the likelihood of finding a target at a particular location. Counterfactuals are employed by cycling images between the PMF heatmap for the previous guidance (an estimation of the features that led to $\widehat{\pi_h^*}$), and the current PMF heatmap (the features that led to π_h^*) at a regular frequency. The numerical probability of finding a target for both the current recommended goal location and the previously recommended goal location is overlaid onto both the prior and current heatmap. This shows explicitly, in numeric form, how the odds have changed to prioritize the current recommendation over the previous one.

Environment-centric Local. This justification uses the same visual representation of alternating between the current and prior PMF as *environment-centric Global*, but instead of showing the entire heatmap, only the heatmap values at the specific goal locations of the current and previous recommendations are shown, alongside the numerical probabilities associated with those two locations.

Policy-centric Global. This justification is conveyed as a natural language counterfactual, focusing on long term rewards. For a multi-target, time-constrained search domain, an example of this justification is "On average, following the new path will result in X targets found overall, compared to the old path at Y targets found." This takes an abstract concept of expected long-horizon reward and maps it to a human understandable sentence. To estimate values X and Y in our partially observable domain, we utilize a heuristic combining the computed odds over the given recommendation with the overall entropy of the PMF, which decreases over time through exploration. This strategy can be employed for any domain that uses a PMF-based goal likelihood formulation.

Policy-centric Local. This justification is also conveyed as a natural language counterfactual, but focused on short-term rewards. For example, our domain uses the form "On average, the new path will take X moves to find a target, compared to the old path at Y moves." The means of generating X and Y in this case is simpler, as the reward can be more accurately estimated over a fixed-horizon recommendation. It is simply a case of mapping abstract reward to human understandable output. Fig. 3 shows how these four justification types were mapped to our treasure-hunt domain.

B. Hypotheses

H1: Objective Hypotheses

H1.a (Compliance): Participants will have higher compliance with recommendations when given policy-based justifications, compared with environment-based justifications and no justification, as policy-based justification utilizes abstraction and framing effects, resulting in a higher level of persuasiveness[44].

H1.b (Performance): Participants will perform better in the game when given policy-based justifications, compared with

environment-based justifications and no justification, as compliance should correlate with performance given the relatively high competence of the recommending system in our domain.
H1.c (Decision-making Time): Participants will take longer to make decisions when given environment-based justifications, compared with policy-based justifications and no justification, as environment-based justification includes more contextual information, which promotes active thinking patterns.

H2: Subjective Hypotheses

H2.a (Mental Load): Participants will report lower mental load when given policy-based justifications, compared with environment-based justifications, since environment-based justifications have more information to process, and compared with no justification, as people tend to report higher workload when interacting with systems behaving inexplicably [59].
H2.b (Trustworthiness): Participants will rate the system as more trustworthy and reliable when given environment-based justifications, compared with policy-based justifications and no justification, as environment-based justification provides more transparency and contextual information, which will result in participants feeling like they understand the decision-making process.
H2.c (Perceived Intelligence): Participants will rate the system as more intelligent when given environment-based justifications, compared with policy-based justifications and no justification, also due to the transparency into the decision-making process provided by environment-based justifications.
H2.d (Justification Interpretability): Participants will rate environment-based justifications as more interpretable, informative, and helpful for decision-making compared to policy-based justifications, due to the extra information provided by environment-based justifications.

VI. Experimental Evaluation

We investigate the preceding hypotheses regarding the effects of different types of justification on participants through an IRB-approved human-subjects study.

A. Experimental Design

We conducted a 5x1 between-subjects experiment using Amazon Mechanical Turk to evaluate the four types of justifications introduced above, alongside a control condition that did not include justifications, in the experimental domain described in Section III (Fig. 1-top). The participants' goal was to explore a maze and find as many buried treasures as they could in a limited number of turns. Participants were assisted in their task by a team of autonomous drone teammates who simultaneously explored the maze and provided constantly-updating recommendations to the human based on their own noisy sensor readings. The VOI-based framework for strategic justification timing described in Section IV determined when justifications should be provided to participants. The type of justifications were determined by experimental condition: 'global policy', 'local policy', 'global environment', 'local environment', or 'no justification' (control).

B. Rules of the Game

Participants played two rounds of the game with the goal of digging up as many of the 25 treasures hidden throughout an 18x27 maze grid as they could in a period of 60 turns. Each turn, participants could choose either to move to any available adjacent grid square, or to dig on the square they currently occupied to earn one treasure if one was located there. A team of AI-controlled drones explored the grid autonomously, moving multiple tiles in a turn and taking noisy treasure-detecting measurements of every tile flown over. These readings were used to update both their PMF and the guidance they provided to the participant. The guidance took the form of a green line with an orange 'X' at the end, indicating where the drones thought the participant should dig next (see Fig. 4), which participants could choose to follow or not. Whenever a justification was triggered by our framework, the prior path recommendation was shown in gray, with the rest of the justification depending on condition (see Fig. 3).

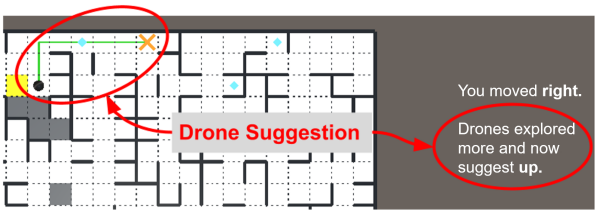

Fig. 4: Drone guidance is shown as a path overlay and a textual representation of the next suggested move.

C. Study Protocol

The experiment was run in several batches with randomly determined condition, using Amazon Mechanical Turk to crowd-source participants. High quality participants were targeted by filtering for high numbers of previously approved tasks on Mechanical Turk, as well as approval percentage. Additionally, on top of the base compensation rate of $3, a bonus of 5¢ per treasure found during the game was paid to further incentivize participant effort towards high performance.

After providing informed consent, participants completed a short pre-experiment demographic survey. After reading the rules of the game, participants completed a short comprehension quiz and played a tutorial level to ensure they understood their objective. Next, participants played the two rounds of the game and completed a post-experiment survey which involved a combination of Likert scale and free response questions.

D. Measurement

The pre-survey collected demographic information about our participants. Out of 104 initial MTurk participants, we removed 13 from data analysis for either failing to locate a single treasure during the game or for repeatedly spending excessive time inactive without inputting a move, indicating lack of understanding of or concentration towards the game, respectively. This left 91 participants (51 males, 37 females, and 3 who did not specify gender) with ages ranging from 23 to 72 years old ($M = 40.99; SD = 11.80$). 39.6% of

participants reported working in a STEM field, and 69.2% of participants reported having received a bachelor's degree or higher. 19 participants each ran the 'global environment' and 'no justification' conditions, 18 each ran the 'global policy' and 'local policy' conditions, and 17 ran the 'local environment' condition.

We collected a number of objective measures from participant gameplay, including:

- *Targets Found*: The total number of treasures discovered.
- *Compliance Rate*: The percentage of moves taken by users that matched the recommendations provided by the system.
- *Compliance Rate During Justification*: The percentage of moves taken by users that matched the recommendations provided by the system, on turns when justifications were provided. Note that in the control condition 'no justification', although justifications are never offered, we still collect this measure by applying the same VOI-timing algorithm but never acting on it.
- *Time Per Move*: The average time taken per move.
- *Time Per Move During Justification*: The average time taken to make decisions when justifications were provided.

TABLE II: Subjective Scale Measure Items.

Trust (Cronbach's α = 0.95)
1. I am confident in the system
2. The system is dependable
3. The system is reliable
4. I can trust the system
Justification Interpretabilty (Cronbach's α = 0.94)
1. I found the justifications to be complete and understandable.
2. I was able to adapt better to the game due to the justifications provided.
3. I found the justifications to be sufficient for making decisions.
4. I found that the justifications were informative during the game.
5. The justifications were useful.
6. I understand why the system used specific information in its justifications.
7. I understood how the system arrives at its answer.
8. I understood the systems reasoning.
9. I could easily follow the justifications to arrive at a decision.
Workload (Cronbach's α = 0.76)
1. How mentally demanding was the game?
2. How hurried or rushed was the pace of the game?
3. How hard did you have to work to accomplish your level of performance?
4. How insecure, discouraged, irritated, stressed, and annoyed were you during the game?
Perceived Intelligence (Cronbach's α = 0.92)
1. System is Competent
2. System is Knowledgeable
3. System is Intelligent
4. System is Sensible
Likert items are coded as 1 (Strongly Disagree) to 7 (Strongly Agree)

For subjective measures, we administered a post-experiment questionnaire to participants after completing the treasure hunt task. The questionnaire was developed using well-established metrics from the fields of robotics and explainable AI, including the Trust in Automation Survey [34], the Interpretability and Decision-Making Surveys for XAI metrics [63, 31, 52], the Stress and Workload (NASA-TLX) [26], and the Perceived Intelligence (Godspeed Questionnaire) [3]. Participants were asked to rate their opinions on the guidance provided by the agent using 7-point Likert-scale items. Based on these questionnaires, we identified four key concepts to validate our hypothesis: *Trust, Justification Interpretability, Workload, and Perceived Intelligence*.

To determine these constructs, we used principal component analysis to extract latent factors from the above mentioned scales and calculated the factor loading matrix using varimax rotation. We identified items that could be combined to create concept scales with a correlation cutoff point of $r \geq 0.6$ to the factor matrix [29] which resulted in the scales presented in table II.

VII. Results

A. Objective Analysis

To test our objective hypotheses, we analyzed the various metrics collected during the game using a one-way analysis of variance (ANOVA) with experimental condition as a fixed effect. Post-hoc tests used Tukey's HSD to control for Type I errors in comparing results across each of the four justification types and the control condition.

Our hypotheses expected between-conditions differences to be more pronounced along the axis of policy-based vs. environment-based features, compared with global vs. local features. Hence, we conducted additional analysis using a one-way ANOVA with bucketed results, comparing policy-based justification vs. environment-based justification vs. no justification. Again, post-hoc significance was determined using Tukey's HSD. The means per condition and per bucket are shown in Tables III and IV below.

	Global Policy	**Local Policy**	**Global Env.**	**Local Env.**	**None**
Compliance Rate*	84.67%^{A}	81.53%	70.65%^{B}	75.48%	70.53%^{B}
Compliance Rate (During Justification)*	56.46%^{A}	54.50%	40.57%^{B}	49.54%	48.52%
Targets Found*	9.28^{A}	8.47$^{A/B}$	7.00$^{B/C}$	7.78	6.32^{C}
Time per Move*	1.30s^{B}	1.40s	2.01s	2.10s^{A}	1.90s
Time per Move (During Justification)*	1.74s^{B}	1.66s^{B}	2.49s	3.39s^{A}	1.85s^{B}

TABLE III: Means for objective measures across all conditions. Measures with ANOVA significance are indicated by *. Post-hoc significance is shown using letters. Individual means denoted by A are significantly higher than B/C or C. Likewise, A/B is significantly higher than C.

The ANOVA revealed significant effects for both overall compliance rate (F(4,86) = 3.98, p = 0.0052), and compliance rate during justification (F(4,86) = 3.09, p = 0.020). Post-hoc analysis for overall compliance rate with Tukey's HSD shows that participants complied significantly more in the 'global policy' condition compared to both the 'no justification' condition (p = 0.019), and the 'global environment' condition (p = 0.020). Post-hoc analysis of compliance rate during justification found a significantly higher compliance in 'global policy' compared to 'global environment' (p = 0.016).

Significance was likewise found in the ANOVA comparing the policy-based, environment-based, and no justification

	Policy Features	Env Features	None
Compliance Rate*	83.14%^{A}	73.00%^{B}	70.53%^{B}
Compliance Rate (During Justification)*	55.51%^{A}	44.93%^{B}	48.52%
Targets Found*	8.89^A	7.38^B	6.32^B
Time per Move*	1.35s^B	2.06s^A	1.90s^A
Time per Move (During Justification)*	1.70s^B	2.93s^A	1.85s^B

TABLE IV: Means for objective measures across the three condition buckets. Measures with ANOVA significance are indicated by *. Individual means denoted by A demonstrated post-hoc significance over means denoted B.

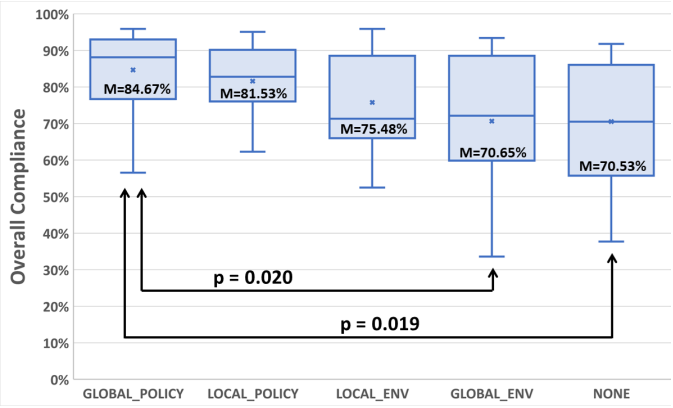

Fig. 5: Compliance rate by condition, with means and post-hoc significance shown.

buckets for both overall compliance rate ($F(2,88) = 7.19$, $p = 0.0013$), and compliance rate during justification ($F(2,88) = 4.41$, $p = 0.015$). Post-hoc analysis showed that overall compliance rate was significantly higher for users with policy-based justifications than those with environment-based justifications ($p = 0.0047$), and those with no justification ($p = 0.0062$). Post-hoc analysis of the compliance rate during justification additionally showed a significant effect for policy-based over environment-based justifications ($p = 0.012$). These results serve to **validate H1.a (compliance)**.

Since our experimental domain was associated with a high degree of robot competence, performance in the game (number of targets found) highly correlated with compliance with the drones' suggestions. Using Pearson's correlation coefficient, we verified this relationship (i.e., the more participants chose to follow the guidance, the better they perform) ($r(91) = 0.77$, $p < 0.0001$). The ANOVA showed a statistically significant effect for number of targets found ($F(4,86) = 4.77$, $p = 0.0016$). Post-hoc analysis showed three significant effects. Participants in 'global policy' found more targets than those in 'no justification' ($p = 0.016$), or in 'global environment' ($p = 0.027$). Additionally, those in 'local policy' found significantly more targets on average compared to 'none' ($p = 0.047$).

The ANOVA per bucket also revealed significance ($F(2,88) = 8.46$, $p = 0.0004$). Post-hoc analysis found that policy-based justifications led to better user performance in the game, compared with both no justification ($p = 0.0005$), and environment-based justifications ($p = 0.018$). These results serve to **validate H1.b (performance)**.

The timing measures, related to the latent measure of participant thinking load, had significant effects both for time per move ($F(4,86) = 3.71$, $p = 0.0078$) and time per move during justification ($F(4,86) = 3.74$, $p = 0.0075$). Post-hoc analysis for time per move showed that participants in the 'local environment' condition took significantly more time to take their moves compared to 'global policy' ($p = 0.030$), but not significantly more time compared to 'local policy' ($p = 0.089$). Additionally, while there was no significant effect for 'global environment' taking longer on average than 'global policy', further exploration may be merited in future work ($p = 0.063$). Post-hoc analysis for time per move during justification showed three significant effects, with 'local environment' taking more time than 'local policy' ($p = 0.016$), 'global policy' ($p = 0.022$), and 'no justification' ($p = 0.033$).

In the bucketed analysis of timing, the ANOVA showed significance in both time per move ($F(2,88) = 7.44$, $p = 0.0010$), and time per move during justification ($F(2,88) = 5.91$, $p = 0.0039$). Post-hoc analysis of time per move showed that, with environment-based justifications, participants took significantly longer than with policy-based justifications ($p = 0.0009$). Interestingly, no justification similarly had a significant effect, taking longer than policy-based justifications ($p = 0.047$). This shows that despite the added cost of attending to justifications, participants were able to take their moves faster on average in the policy-based justification conditions. Similarly, post-hoc analysis of time per move during justification showed that environment-based justifications took significantly higher time than both policy-based justifications ($p = 0.0049$), and no justifications ($p = 0.050$). These results serve to **validate H1.c (decision-making time)**.

B. Subjective Analysis

We conducted similar analysis to test our subjective hypotheses, running one-way ANOVAs fixed by both experimental condition, as well as bucketed by the feature class seen during justification (policy-based, environment-based, or no justification). Post-hoc significance was determined using Tukey's HSD. In the case of the scale for justification interpretability, the Likert-scale questions asked referred specifically to justifications, so was limited only to the four experimental conditions that possessed justifications, excluding the control.

Of the 91 participants with usable gameplay data, an additional five failed basic attention-check questions in the survey. Post-hoc analysis of survey responses showed six further outliers, with significantly lower internal consistency among related survey question answers than other participants, appearing more like random clicking than coherent responses. Removal of those 11 participants left us with the surveys of 80 participants for subjective analysis.

There were no statistically significant differences on the *Workload* scale, either in the ANOVA with experimental

	Global Policy	Local Policy	Global Env.	Local Env.	None
Workload	3.40	3.67	4.05	3.63	4.24
Trust	4.15	3.94	5.23	4.80	4.87
Perceived Intelligence	4.59	4.88	5.73	5.16	5.27
Justification Interpretability*	4.32^B	4.24^B	5.40^A	4.96	N/A

TABLE V: Means for subjective measures across all conditions. Measures with ANOVA significance are indicated by *. Individual means denoted by A demonstrated post-hoc significance over means denoted B.

	Policy Features	Env Features	None
Workload	3.53	3.85	4.24
Trust*	4.05^B	5.03^A	4.87
Perceived Intelligence*	4.73^B	5.47^A	5.27
Justification Interpretability*	4.28^B	5.20^A	N/A

TABLE VI: Means for subjective measures across all conditions. Measures with ANOVA significance (or Student's t-test significance, in the case of Justification Interpretability) are indicated by *. Individual means denoted by A demonstrated post-hoc significance over means denoted B.

condition as its fixed effect or between the bucketed classes of policy-based, environment-based, and no justification. Therefore, the hypothesis **H2.a (mental load) is inconclusive**.

The condition-wise ANOVA of the *Trust* scale also did not reveal a significant effect ($F(4,75) = 2.33$, $p = 0.064$), but the bucketed ANOVA for *Trust* did reveal significance ($F(2,77) = 4.29$, $p = 0.017$). Post-hoc analysis with Tukey's HSD revealed that environment-based justifications were rated as significantly more trustworthy than policy-based justifications ($p = 0.019$). However, no effect was found between environment-based justification conditions and no justification, meaning this result serves to **partially validate H2.b (trustworthiness)**.

Likewise, while the per condition ANOVA of the *Perceived Intelligence* scale was not significant ($F(4,75) = 2.23$, $p = 0.073$), the feature-class bucketed ANOVA for *Perceived Intelligence* was ($F(2,77) = 3.30$, $p = 0.042$). Post-hoc analysis showed that the drone teammates using environment-based justifications were rated as significantly more intelligent than the drone teammates using policy-based justifications ($p = 0.038$). Again, no effect was found between environment-based conditions and no justification, meaning this result serves to **partially validate H2.c (perceived intelligence)**.

Lastly among the subjective scales, the ANOVA for the *Justification Interpretability* scale did reveal significance when

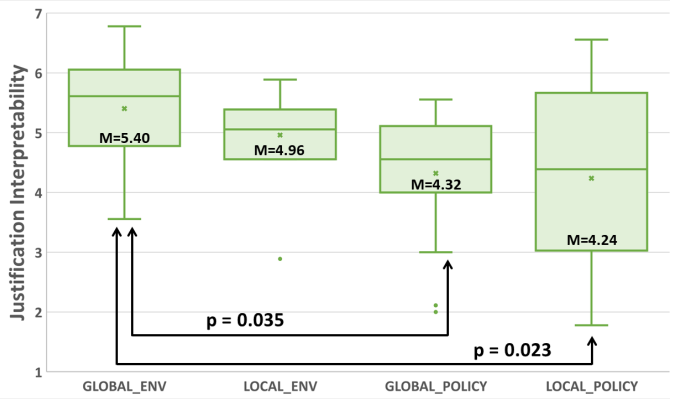

Fig. 6: Rated interpretability of justifications by condition, with means and post-hoc significance shown.

fixed by experimental condition ($F(3,59) = 3.94$, $p = 0.013$). Post-hoc analysis revealed that the justifications in the 'global environment' condition were rated as significantly more interpretable and informative when compared to the justifications from both the 'local policy' condition ($p = 0.023$), and the 'global policy' condition ($p = 0.035$).

There was an additional significant effect for the data bucketed by feature class for the *Justification Interpretability* scale. Since this scale specifically compares justifications, the 'no justification' bucket is excluded from analysis, and the data is compared using a simple one-tailed t test, where the justifications from environment-based justification conditions are rated as significantly more interpretable compared to justifications from policy-based justification conditions ($t(61) = -3.35$, $p = 0.0007$). These results serve to **validate H2.d (justification interpretability)**.

VIII. Recommendations & Potential Applications

A. Recommendations for Justification Design

In this section, we summarize the main findings and implications drawn from the results of our user study on the utility of justification in human-robot interaction.

1) High Robot Competence or Low Human Competence: Use Policy-based Justifications: Policy features are highly comprehensible to human teammates, as the information is packaged such that users can compare the end results of the robot's decision making. The information is highly abstract, and is framed taking the human teammate's own utility into account. There is little room to think critically about or question the accuracy of policy-based counterfactual justifications, which resulted in a high level of persuasiveness in our study (we saw that policy-based justifications led to significantly higher compliance when compared with environment-based or no justifications). In our user study with highly competent robot teammates, participants were more successful in accomplishing their task when presented with this style of low transparency, easily comprehensible justification.

It is important to note that if the robotic agent were not giving competent recommendations, participants would likely

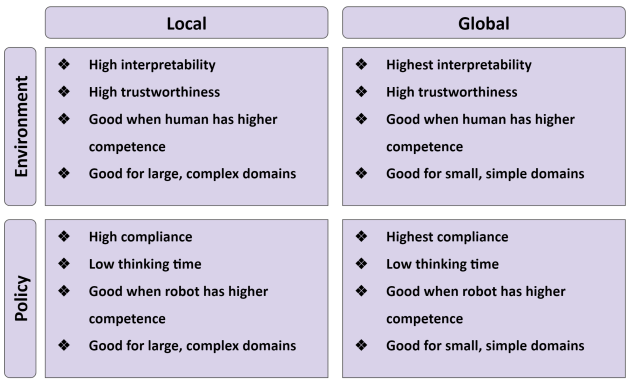

Fig. 7: A taxonomy of the usefulness of each justification type.

have performed significantly worse due to their over-reliance on a low-quality decision support system. Policy-based justification could result in over-reliance and dependence on the system, causing passive thinking patterns [35] where the human cedes effective control of decision-making entirely to the robot agent. In cases of low robot competence, this would lead to a large number of Type I errors where users accept low-quality advice from the system [21, 14]

Therefore, during human-robot teaming scenarios or domains where you would expect the quality of robotic guidance to be fairly high relative to a human operating by themselves, policy-based justification should be used, increasing human teammate compliance, making them a more predictable member of a multi-agent team. This would significantly improve the planning system's ability to optimize over all agents, since the innate uncertainty associated with accounting for human decision making would be greatly reduced [13, 39]. Policy-based justification can also be suitable when the human needs to make snap decisions in time-critical situations.

2) Low Robot Competence or High Human Competence: Use Environment-based Justification: Environment-based features provide highly interpretable, highly contextual information, and are well-suited for representing uncertainty. They push human teammates towards a more active thinking pattern, which is more analytical, deliberate, and rational [35]. Humans tend to view this type of justification as more of a tool, compared with the more abstracted policy-based justifications. This can lead to better-informed decision making and more successful adaptation to uncertain situations. In our study, we observed that environment-based justifications during changes of recommendation were associated with significantly more thinking time than policy-based or no justifications. What's more, participants rated robotic agents using environment-based justifications as the most trustworthy, and environment-based justifications themselves as the most informative, interpretable, and helpful for their decision-making process.

This added transparency and increased information content comes at the cost of being more demanding and time-consuming to parse, leading to slower decisions. Additionally, environment features are able to be interpreted in any number of ways by different human agents, which often leads to highly variable, independent human behavior [59]. This leads to a significantly lower compliance rate when compared with policy-based justifications. If environment-based justifications were deployed in a domain with a high relative competence of robot-provided guidance, there would be a large number of Type II errors made, whenever users reject the high-quality advice of the robot. Therefore, in scenarios where the human teammate brings expertise in their decision-making that is hard to match with the automated guidance of a collaborative robot, environment-based justifications are more appropriate.

Focusing on the other axis of our 2x2 justification characterization, in our study we generally found that the use of global features outperformed the local features on the respective measures that policy-based and environment-based justifications excelled at. For instance, 'global policy' had the highest user compliance rate and performance, and 'global environment' had the highest perceived interpretability. We posit that this is likely related to the short-term nature of the interaction in our evaluation domain. In longer lasting, more complex domains, local features may prove may beneficial, as they can help prevent the human teammate from being overwhelmed by excess information. More research is needed to confirm this. We summarize the characteristics and suitable use cases of each justification type in Fig. 7.

B. Potential Application: AR-based Spatial Navigation

To illustrate the application of these synthesized justification design principles, we present a concept of how they might by implemented in a real-world decision support system embedded in an augmented reality (AR) interface (similar to Tabrez et al. [59]). Since our framework and results are drawn from a partially observable, multi-goal search task, we designed this interface for domains that share these characteristics, such as search and rescue, radiological device recovery, or explosive ordnance disposal. However, since the features tested were derived from general xAI principles, it is likely that the taxonomy presented in Fig. 7 is more broadly applicable to a wide range of human-robot collaborative tasks, though further research is needed to confirm this.

Humans using this interface explore an environment searching for hidden targets. Meanwhile, a drone teammate conducts its own exploration of the environment, using its sensors to update its model of where it believes the hidden targets are likely to be. The drone continually provides navigation guidance to the human, aiding them in the task of locating as many targets as possible in a limited amount of time. Whenever justification is triggered by a significant change in guidance, one of two justification modules is chosen, depending on the drone's current confidence in the quality of that guidance.

AR-based Policy Justification. In regions of high drone confidence, a policy justification is triggered (Fig. 8 Top). The AR interface renders the current guidance in the form of a colored arrow and pin directly overlaid onto the environment, telling the human where the drone thinks they should go and search next. The guidance from the prior time step is rendered as a gray arrow and pin. In addition to these

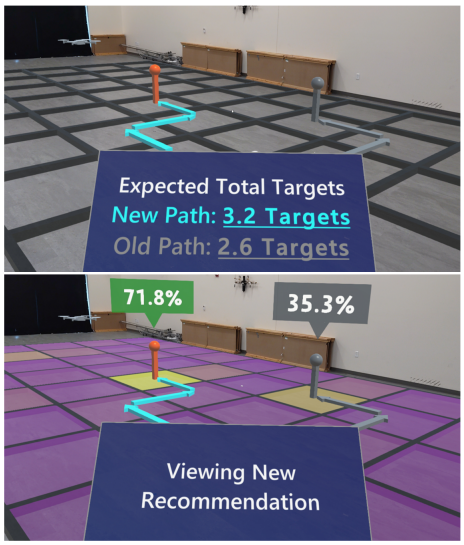

Fig. 8: Top: AR-based policy justification. Bottom: AR-based environment justification.

paths, a counterfactual natural language description is provided as justification on the user's AR-based menu, showing the difference in expected utility of taking the new path in contrast to the old path.

AR-based Environment Justification. In regions of low drone confidence, an environment justification is triggered (Fig. 8 Bottom). In addition to rendering the current and previous paths as seen in the policy justification, the AR interface renders the drone's current PMF as a heatmap overlaid onto the environment, using a gradient from purple to yellow to represent low and high chances of finding a target, respectively. Two AR-based pins are rendered over the current and prior targets, showcasing the local PMF values at each location. Users are able to view the PMF and pins from the prior timestep to visualize how the environment features changed to lead to a changed recommendation, providing a justification for taking the new path as opposed to the old path.

The task in this implementation has similar dynamics to the treasure hunt game, though lifted into a 3D, real world domain. Although the interface pictured in Fig. 8 is shown at the scale of a large room, the same type of visualization could be spatially expanded to large outdoor environments to serve as a viable interface for real-world drone assisted target-finding tasks.

IX. Conclusion

In this work, we highlighted the value of strategic timing for robot-provided explanations that serve as justifications during instances of mismatched expectations in the context of decision-support for human-robot teaming (e.g., when an agent's recommendation is unexpected or confusing). A justification provided in this context aims to convince the human teammate of the utility of the previously difficult-to-interpret recommendations. Our work contributes answers toward two fundamental questions at the intersection of explainable AI and human-robot teaming: 1) When are justifications most impactful and useful? And 2) What information should be presented in those justifications to improve human teammate decision-making and behavior?

We propose a novel value of information-based framework to determine when a decision-support system should provide justifications to a human collaborator, such that a balance is struck between informativeness, and avoiding habituation and excess cognitive load. We validated the proposed framework through an expert-feedback case study, demonstrating the usefulness of justifications when they are timed appropriately. We also present a characterization of four types of counterfactually generated justification, drawing from a taxonomy established in explainable AI literature: **global policy**, **local policy**, **global environment**, and **local environment**. The justification types were evaluated in an online human subjects study ($n = 91$) involving a collaborative, partially observable search task alongside robot teammates.

We show that robots providing policy-based justification led to higher compliance and faster decision-making. We additionally show, in contrast, that robots providing environment-based justification led to higher subjective ratings of interpretability, intelligence, and trustworthiness of the robot teammates.

Based on our experimental findings, we offer actionable recommendations for operationalizing these results into decision-support systems that prioritize explainability and foster appropriate trust and reliability. We additionally demonstrate how these sythesized design principles can be applied to a real-world decision-support system with a concept augmented-reality interface. Justifications should be user-centric, taking into consideration the relative competence of human and robotic agents, the user's expectations of the robot, and how different types of justification can influence user thinking patterns and performance.

Acknowledgments

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

References

[1] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1168–1176, 2018.

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81, 2009.

[4] Maya Cakmak and Andrea L Thomaz. Designing robot learners that ask good questions. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 17–24. IEEE, 2012.

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

[6] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the international conference on automated planning and scheduling*, volume 29, pages 86–96, 2019.

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

[8] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The emerging landscape of explainable automated planning & decision making. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 2021.

[9] 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.

[10] Nicholas Conlon, Daniel Szafir, and Nisar Ahmed. Investigating the effects of robot proficiency self-assessment on trust and performance. *arXiv preprint arXiv:2203.10407*, 2022.

[11] 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.

[12] Francisco Cruz, Charlotte Young, Richard Dazeley, and Peter Vamplew. Evaluating human-like explanations for robot actions in reinforcement learning scenarios. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 894–901. IEEE, 2022.

[13] Abhinav Dahiya, Alexander M Aroyo, Kerstin Dautenhahn, and Stephen L Smith. A survey of multi-agent human–robot interaction systems. *Robotics and Autonomous Systems*, 161:104335, 2023.

[14] Stephen R Dixon and Christopher D Wickens. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors*, 48(3):474–486, 2006.

[15] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.

[16] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[17] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[18] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[19] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274, 2019.

[20] John R Frost. *The theory of search: a simplified explanation*. Soza Limited, 1997.

[21] Matthew Gombolay, Xi Jessie Yang, Bradley Hayes, Nicole Seo, Zixi Liu, Samir Wadhwania, Tania Yu, Neel Shah, Toni Golen, and Julie Shah. Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research*, 37(10):1300–1316, 2018.

[22] 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.

[23] Kalanit Grill-Spector, Richard Henson, and Alex Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006.

[24] Matthew Grizzard, Ron Tamborini, John L Sherry, René Weber, Sujay Prabhu, Lindsay Hahn, and Patrick Idzik. The thrill is gone, but you might not know: Habituation and generalization of biophysiological and self-reported arousal responses to video games. *Communication Monographs*, 82(1):64–87, 2015.

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

[26] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[27] Bradley Hayes and Brian Scassellati. Challenges in shared-environment human-robot collaboration. *learning*, 8(9).

[28] 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.

[29] Guy Hoffman and Xuan Zhao. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–31, 2020.

[30] Guy Hoffman, Maya Cakmak, and Crystal Chao. Timing in human-robot interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 509–510, 2014.

[31] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[32] Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1): 22–26, 1966.

[33] Jeffrey L Jenkins, Bonnie Brinton Anderson, Anthony Vance, C Brock Kirwan, and David Eargle. More harm than good? how messages that interrupt can make us vulnerable. *Information Systems Research*, 27(4):880–896, 2016.

[34] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.

[35] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

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

[37] Bing Cai Kok and Harold Soh. Trust in robots: Challenges and opportunities. *Current Robotics Reports*, 1: 297–309, 2020.

[38] Anagha Kulkarni, Yantian Zha, Tathagata Chakraborti, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. Explicable planning as minimizing distance from expected behavior. In *AAMAS Conference proceedings*, 2019.

[39] Shih-Yun Lo, Elaine Schaertl Short, and Andrea L Thomaz. Planning with partner uncertainty modeling for efficient information revealing in teamwork. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 319–327, 2020.

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

[41] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[42] Ruikun Luo, Na Du, and X Jessie Yang. Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human–Computer Interaction*, 38(18-20):1962–1971, 2022.

[43] Walter Mischel. *The marshmallow test: Why self-control is the engine of success*. Little, Brown New York, 2015.

[44] Scott Plous. *The psychology of judgment and decision making*. Mcgraw-Hill Book Company, 1993.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[46] Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, 2019.

[47] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. *Active preference-based learning of reward functions*. 2017.

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

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[50] Raymond Sheh. Explainable artificial intelligence requirements for safe, intelligent robots. In *2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*, pages 382–387. IEEE, 2021.

[51] Thomas B Sheridan. Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532, 2016.

[52] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, pages 1–15, 2022.

[53] Herbert A Simon. Bounded rationality. *Utility and probability*, pages 15–18, 1990.

[54] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Balancing explicability and explanation in human-aware planning. In *2017 AAAI Fall Symposium*, pages 61–68. AI Access Foundation, 2017.

[55] Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. Balancing communication and behavior. In *Explainable Human-AI Interaction: A Planning Perspective*, pages 95–105. Springer, 2022.

[56] 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.

[57] 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.

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

[59] 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*, pages 1256–1264, 2022.

[60] 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.

[61] Richard F Thompson and William A Spencer. Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological review*, 73(1):16, 1966.

[62] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[63] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(3): 1–24, 2021.

[64] Michał Wysokiński, Robert Marcjan, and Jacek Dajda. Decision support software for search & rescue operations. *Procedia Computer Science*, 35:776–785, 2014.

[65] Lu Yadong and Zhou Ya. Optimal search and rescue model: Updating probability density map of debris location by bayesian method. *International Journal of Statistical Distributions and Applications*, 1(1):12, 2015.

[66] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 408–416, 2017.