2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-7281-6212-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/IROS45743.2020.9341792

Online Explanation Generation for Planning Tasks in Human-Robot Teaming

Mehrdad Zakershahrak, Ze Gong, Nikhillesh Sadassivam and Yu Zhang

***Abstract*— As AI becomes an integral part of our lives, the development of explainable AI, embodied in the decision-making process of an AI or robotic agent, becomes imperative. For a robotic teammate, the ability to generate explanations to justify its behavior is one of the key requirements of explainable agency. Prior work on explanation generation has been focused on supporting the rationale behind the robot's decision or behavior. These approaches, however, fail to consider the mental demand for understanding the received explanation. In other words, the human teammate is expected to understand an explanation no matter how much information is presented. In this work, we argue that explanations, especially those of a complex nature, should be made in an online fashion during the execution, which helps spread out the information to be explained and thus reduce the mental workload of humans in highly cognitive demanding tasks. However, a challenge here is that the different parts of an explanation may be dependent on each other, which must be taken into account when generating online explanations. To this end, a general formulation of online explanation generation is presented with three variations satisfying different *"online"* properties. The new explanation generation methods are based on a model reconciliation setting introduced in our prior work. We evaluated our methods both with human subjects in a simulated rover domain, using NASA Task Load Index (TLX), and synthetically with ten different problems across two standard IPC domains. Results strongly suggest that our methods generate explanations that are perceived as less cognitively demanding and much preferred over the baselines and are computationally efficient.**

I. INTRODUCTION

As intelligent robots become more prevalent in our lives, the interaction of such AI agents with humans becomes more frequent and essential. One of the most important aspects of human-robot interaction is for the robotic agent to provide explanations to support the rationale behind its decision or behavior [1]. An explanation provides justifications for the robot, which helps the human maintain trust of the robotic peer as well as a shared situation awareness [2], [3]. Prior work on explanation generation, however, often ignores the underlying requirements of the human recipient to understand an explanation [4], [5], [6]. A good explanation should be generated in a lucid fashion from the recipient's perspective [7], [8], so that it is understood.

To address this problem, a key consideration is that the human recipient (*explainee*) may interpret an explanation differently from the robot (*explainer*) due to a different understanding of the domain. In our prior work [7], we refer to such differences as *model differences*. The robotic agent, as a result, must ensure that the explanation makes sense in the human's model, which generates the human's expectation of the robot, so that the robot's behavior matches with the human's expectation. An explanation can then be considered as a request to change the human's model to reduce the model differences so that the robot's behavior is consistent with the updated human model. The decision-making process (including explanation generation discussed herein) in the presence of such model differences is more generally referred to as *model reconciliation* [7], [9].

One remaining challenge, however, is the consideration of the mental demand of the human to understand an explanation. In most prior work on explanation generation, the human is expected to understand an explanation regardless of how much information it contains. Little discussion has been given on the ways of presenting such information. In this work, we argue that explanations, especially complex ones, should be provided in an online fashion, such that each explanation is broken into multiple parts, which are then communicated separately and intertwined with plan execution. Communicating an explanation in such a manner is expected to result in less mental workload for cognitively demanding tasks since the information is spread out so that the interpretation process becomes incremental, which is known to benefit understanding [10]. One of the main challenges here is that the different parts of an explanation could be dependent on each other, which must be taken into account when generating online explanations. Our online explanation generation process spreads out the information while ensuring that the different parts do not introduce cognitive dissonance so that they are always perceived in a cohesive fashion.

A. Motivating Example

Let us illustrate the motivation of online explanation generation via a familiar situation involving a daily routine. Mark works at a company and has a voice assistant helping him get ready for work everyday. Usually Mark wakes up, drinks a freshly made coffee, enjoys a filling breakfast, dresses for work, and then drives to work. However, today, Mark has a meeting scheduled in the early morning so Mark needs to arrive at work earlier, a presentation at near lunch time, his car is broken, and there is no coffee beans for fresh coffee. The voice assistant knows that Mark must be reminded of these changes. However, explaining all the changes at the same time may result in unnecessary strain on Mark. In contrast, the robot first suggests Mark to prepare an instant coffee,, explaining that there is no coffee beans

Mehrdad Zakershahrak, Ze Gong, Nikhillesh Sadassivam and Yu Zhang are with the School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ. {mzakersh,zgong11,nsadassi,yzhan442}@asu.edu

left. As Mark is enjoying his coffee, it reminds him to cook a light breakfast since there is an important meeting scheduled early today. As Mark is enjoying his breakfast, the voice assistant advises Mark to prepare a lunch box, since there is a presentation at near lunch time so that Mark may not have time to eat outside. After Mark is done with the lunch box, it asks Mark to call the taxi company since his car is broken. After breakfast, the assistant mentions that Mark needs to dress up today for work. When the taxi arrives, the voice assistant asks Mark to take the lunch box. Comparing the voice assistant's strategy to share all of the information at the beginning, we can see that conveying the information in an online fashion is more cognitively friendly (i.e., involving less information at a time) and hence helps with reducing strain and cognitive load [11], [12]. These effects are highly desirable for tasks that are cognitively demanding for humans.

In this paper, we develop a general formulation of online explanation generation by breaking an explanation up into multiple parts to be communicated at different times during plan execution. We develop three variations of online explanation generation methods with each satisfying different "online" requirements. In the first method, the focus is for a robot to explain only plan prefixes. This is in contrast to prior offline methods where the entire plan must be explained, which allows us to break an explanation up into multiple parts with each explaining only a part of the plan. We use a model search method to ensure that the earlier parts communicated do not affect the latter parts of an explanation. In the second method, we further relax the online requirement by requiring only the very next action to be explained (if needed). The assumption here is that the actions already occurred do not affect the understanding of the robot's future actions, which holds in situations where each action is viewed independently or the human has a short cognitive span (such as in highly demanding tasks). In the third method, we relax the assumption of the uniqueness of the human's interpretation of a plan and the robot is only required to explain with respect to any such interpretation. A compilation method is developed that converts this problem into one that requires solving two planning problems. Our methods are evaluated both synthetically and with human subjects in standard planning domains. Results strongly suggest that our methods not only generate explanations that are perceived as less cognitively demanding and much preferred over the baselines but also are computationally efficient.

II. Related Work

The advancement of AI and its numerous applications have provided astounding benefits in many areas such as transportation, medicine, finance, education, and entertainment. And yet AI agents have thus far been limited in their ability to operate as a teammate. To be considered a teammate, an AI agent must not only achieve a given task, but also provide a level of transparency about itself to other members of the team [3]. One way to achieve this is to enable AI agents to be self-explanatory in their behaviors. Recently, the explainable AI paradigm [13] rises as one essential constituent of AI systems. Explainable AI maintains a shared situation awareness by facilitating the human's understanding of the AI agent, which also improves the human's trust.

The effectiveness of explainable agency [14] depends on the agent's ability to model the human's interpretation of its behavior. While there exists prior work that focuses on aligning the values [15] or goals [16], [17], the interpretation also depends on the domain model [18]. This means that an explainable AI agent must not only model the domain, but also the human's interpretation of the domain [19], which may be quite different. This interpretation model of the human enables the AI agent to infer about the human's expectation of itself. Using such a model, an agent can generate explicable plans [9], [20], [21], assistive actions [22], etc., to facilitate fluent human-robot interactions. In these methods, the AI agent substitutes a cost metric with a new metric that simultaneously considers the cost and a distance metric between the robot's behavior and the human's expectation of it. Optimizing this new metric often leads to a trade-off between the plan cost and plan interpretability.

Another way to use the human's interpretation of the domain model requires explicit communication, which has the benefit of maintaining cost optimality. The model can be used to infer about which actions in an optimal robot plan are likely to introduce misinterpretations. In some cases, simply providing the future context for those actions is sufficient [23] to make them interpretable. Methods for analyzing the domain to identify the "*causes*" of the robot's plan (or its failures) have been studied before [4], [5], [6]. These methods however assume no differences between the robot's and human's models. More recently, research work has been proposed to specifically address this issue by considering their differences and generating explanations to reduce them [7], [8]. However, all research above has been focused on generating the "*right*" explanation while ignoring the cognitive requirement of the human for understanding the explanation. In our prior work, we have studied how the ordering for presenting the information of an explanation may influence its interpretation [24]. In this work, we further argue that an explanation should be made in an online fashion for cognitively demanding tasks.

III. Explanation Generation as Model Reconciliation

We consider the explanation generation problem in a model reconciliation setting first introduced in our recent work [7]. The reason for this choice is that it represents a more general setting for explanation generation than those used in the previous work, which considers both the robot's (*explainer*) and human's (*explainee*) models as discussed in the related work. An illustration of the model reconciliation setting is presented in Fig. 1. Next, we provide a brief review of the formulations used in our setting.

Model reconciliation defines a planning setting. A planning problem is defined as a tuple (F, A, I, G) using PDDL [25], which is similar to STRIPS [26]. $M = (F, A)$

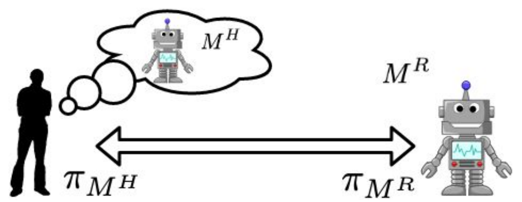

Fig. 1: The model reconciliation setting first introduced in [7]. M^R represents the robot's model and M^H the human's interpretation model of the robot's behavior. Using M^H, the human obtains his expectation of the robot's behavior π_{M^H}. Whenever that is inconsistent with the robot's actual behavior π_{M^R} (generated by M^R), the robot explains by generating an explanation to reconcile the two models.

is also referred to as the *model* in this work, where F is the set of predicates used to specify the state and A the set of actions used to update the state. Actions are associated with a set of preconditions, add and delete effects. I, G are the initial and goal states, respectively.

Definition 1 (Model Reconciliation [7]): A model reconciliation setting is a tuple $(\pi^*_{I,G}, \langle M^R, M^H\rangle)$ $(M^R \neq M^H)$ under a given I, G, where $\pi^*_{I,G}$ corresponds to π_{M^R} in Fig. 1 and represents the robot's behavior (plan) to be explained.

Assuming rational agents, the $\pi^*_{I,G}$ above must satisfy $cost(\pi^*_{I,G}, M^R) = cost^*_{M^R}(I, G)$, where $cost(\pi, M)$ returns the cost of a plan π under the model M, and $cost^*_M(I, G)$ returns the cost of the optimal plan for the given initial and goal states under M. In other words, the robot's plan to be explained must be optimal under M^R. It is assumed that the human obtains his expectation of the robot using M^H. Hence, when the robot's behavior does not match with the human's expectation, explanations must be made. The goal of model reconciliation is to make the robot's plan $\pi^*_{I,G}$ also interpretable under the human's model M^H (i.e., generable by M^H) by reducing the differences between M^H and M^R.

To define model differences, a mapping function Γ was defined in [7] to convert a planning problem into a set of features that fully specify the given problem. For simplicity, we modify the function here to remove the consideration of differences in the initial and goal states. As such, Γ maps any planning problem from its model space $\mathcal{M}$ to the power set of its feature space $\mathcal{F}$ (i.e., $\Gamma : \mathcal{M} \longmapsto 2^{\mathcal{F}}$) as follows:

$$\tau(f) = \begin{cases} a - has - precondition - f, & \text{if } f \in pre(a), a \in A. \\ a - has - add - effect - f, & \text{if } f \in eff^+(a), a \in A. \\ a - has - del - effect - f, & \text{if } f \in eff^-(a), a \in A. \\ a - has - cost - f, & \text{if } f = c_a, a \in A. \end{cases}$$

$$\Gamma(M = (F, A)) = \{\tau(f) | \bigcup_{a \in A} \{f | f \in \{c_a\} \cup pre(a) \cup eff^+(a) \cup eff^-(a)\}\}$$

Definition 2 (Explanation Generation [7]): An explanation in a model reconciliation setting $(\pi^*_{I,G}, \langle M^R, M^H\rangle)$, is a set of unit feature changes Δ to M^H such that 1) $\Delta = \Gamma(\widehat{M^H}) \backslash \Gamma(M^H) \subseteq \Gamma(M^R)$, and 2) $cost(\pi^*_{I,G}, \widehat{M^H}) - cost^*_{\widehat{M^H}}(I, G) < cost(\pi^*_{I,G}, M^H) - cost^*_{M^H}(I, G)$, where $\widehat{M^H}$ is the model after the changes.

An explanation hence reconciles M^R and M^H by reducing their differences and making the cost difference between the human's expected plan $cost^*_{\widehat{M^H}}(I, G)$ and the robot's plan $cost(\pi^*_{I,G}, \widehat{M^H})$ smaller after the model updates. When the cost difference becomes 0, the robot's plan becomes optimal (and hence aligned with that) in the human's model.

Definition 3 (Complete Explanation [7]): An explanation is complete if it satisfies $cost(\pi^*_{I,G}, \widehat{M^H}) = cost^*_{\widehat{M^H}}(I, G)$.

A minimal complete explanation (MCE) [7] is defined as a complete explanation that contains the minimum number of unit feature changes.

IV. Online Explanation Generation (OEG)

Prior work on explanation generation (including [7]) focuses on providing the rationale behind the robot's decision making. It is often assumed that the explanation is provided in its entirety before the execution. As such, the cognitive requirement of the human for understanding the explanation is largely ignored: when complex explanations are involved (such as in cognitively demanding tasks), communicating all information at the beginning becomes impractical. In such cases, it is desirable to communicate explanations in an incremental fashion. In this paper, we introduce *online explanation generation* to address the above issue. The key idea here is to break up an explanation into multiple parts while ensuring consistency for interpretation, and communicate them separately during plan execution. Each part of an explanation is referred to as a *sub-explanation*. A key observation that allows us to break up an explanation is that only a part of the robot's plan needs to be explained by each sub-explanation given at a specific time step. Next, we discuss three variations of OEG methods.

A. OEG for Matching Plan Prefix (OEG-PP)

In this variation, each sub-explanation is required to explain a prefix of the robot's plan, such that it is consistent with the prefix of the human's expectation of the robot's plan. The sub-explanations are made incrementally in the sense that each sub-explanation, when combined with the previous ones, explains a longer prefix of the robot's plan. The implication here is that the human's expected plan after all the sub-explanations will necessarily be the same as the robot's plan, which is the longest prefix of itself.

Definition 4 (OEG-PP): An online explanation for matching plan prefix is a set of sub-explanations in the form of $\langle e_k, t_k\rangle$, where e_k represents the set of unit feature changes to be made as the kth sub-explanation *before* executing step t_k (the t_kth action) of the plan, such that the following holds:

$$\forall k > 0, \mathit{Prefix}(\pi^*_{I,G}, t_k - 1) = \mathit{Prefix}(\pi^H_{E_{k-1}}, t_k - 1)$$
$$s.t. \quad \Gamma(M^H_{E_{k-1}}) = \Gamma(M^H) \cup E_{k-1},$$
$$E_{k-1} = \bigcup_{1}^{k-1} e_i, \text{ and } E_{k-1} \subseteq \Gamma(M^R) \quad (1)$$

where *Prefix*(π, t) returns the prefix of a plan π up to step t (inclusive). E_k represents the union of all sub-explanations up to the kth sub-explanation and $\pi^H_{E_k}$ the optimal plan created under $M^H_{E_k}$, which denotes M^H after incorporating all the changes from e_1 to e_k. More intuitively, at any step $k-1$, the corresponding sub-explanation e_{k-1} is only responsible to explain the actions from t_{k-1} and onward until $t_k - 1$ in the robot's plan.

To generate each $\langle e_k, t_k \rangle$, the search process must consider how the sequence of model changes as a result of each sub-explanation would result in the change of the human's expectation. This allows us to convert the problem of online explanation generation to the problem of model space search as in [7]. The challenge here is that the model changes are not independent, i.e., future sub-explanations may have violated the condition in Eq. (1) for the earlier sub-explanations. In such cases, an online explanation may become undesirable since the human may question the robot's earlier actions at a later stage, even though they appeared reasonable. This situation would introduce cognitive dissonance that may affect the human's understanding of the robot's plan.

To address this issue, it must be ensured that the model changes in the sub-explanations e_k and onward, would not change the plan prefix that is already established up to plan step $t_k - 1$. This can be achieved by searching backward from M^R to M^H. More specifically, given the model reconciliation setting for an explanation generation problem $(\pi^*_{I,G}, \langle M^R, M^H \rangle)$, the following process can be performed recursively to determine each sub-explanation. First, we compute the human's expected plan using M^H, which is denoted as π_H. Denote the index of the first action where $\pi^*_{I,G}$ and π_H differ as t_1, which is the timing for the first sub-explanation. To determine e_1, our search starts from M^R. It finds the largest set of model changes to M^R, denoted as $\overline{e_1}$, such that *Prefix*$(\pi^*_{I,G}, t_1)$ = *Prefix*(π^H_M, t_1) under any M that lies in between $M^R \setminus \overline{e_1}$ and M^R (i.e., $\Gamma(M^R) \setminus \overline{e_1} \subseteq \Gamma(M) \subseteq M^R$). In this way, we guarantee that no change in $\overline{e_1}$ can violate the condition in Eq. (1) once all the feature changes in e_1 are explained, which is exactly what we strive for! e_1 is then computed as the complement of $\overline{e_1}$, or $e_1 = \Gamma(M^R) \setminus (\Gamma(M^H) \cup \overline{e_1})$. Now that we have found $\langle e_1, t_1 \rangle$, we can set M^H to be $\Gamma(M^R) \setminus \overline{e_1}$, or equivalently $\Gamma(M^H) \cup e_1$ (i.e., $M^H_{E_1}$), to determine the next sub-explanation in a recursive manner. The recursion stops when the human's expected plan under $\Gamma(M^H) \cup \bigcup_{i=1}^{k} e_i$ (i.e., $M^H_{E_k}$) matches with $\pi^*_{I,G}$ for the first time, where e_k becomes the last sub-explanation.

The model space search in OEG-PP for determining the kth sub-explanation is illustrated in Fig. 2. In practice, this search is computationally expensive. Hence, we implement an approximate method that searches forward from $M^H_{E_{k-1}}$ for the kth sub-explanation. The search is stopped when the smallest e_k that satisfies *Prefix*$(\pi^*_{I,G}, t_k)$ = *Prefix*$(\pi^H_{E_k}, t_k)$ is found. This approach is more efficient but comes with the cost that no guarantee can be made regarding the latter sub-explanations–they may introduce violations to the condition in Eq. (1) for the earlier steps. If this happens, we backtrack. The implication here is that this method can no longer be used as an *online planning* method (i.e., computing the e_k's online): even though the sub-explanations are communicated online, they must be created offline.

B. OEG for Matching Next Action (OEG-NA)

In this variation, we relax the requirement in Eq. (1) by requiring only the very next action to be interpretable at any step. The assumption here is that the human would not evaluate the robot's behavior retrospectively (or that its influence is minimal), which is reasonable in cognitively demanding tasks where humans must focus more on the current situation due to a very limited cognitive span in such cases [27]. It is also worth noting that OEG-PP and OEG-NA represent the two ends of the spectrum for online explanation generation where OEG-PP considers all actions occurred previously while OEG-NA ignores them all. It is expected that some method in between may work the best. Such analysis will be performed in our future work.

Definition 5 (OEG-NA): An online explanation for matching next action is a set of sub-explanations in the form of $\langle e_k, t_k \rangle$ such that the following is satisfied:

$$\begin{aligned} &\forall k > 0, \pi^*_{I,G}[t_{k-1} : t_k - 1] = \pi^H_{E_{k-1}}[t_{k-1} : t_k - 1] \\ &\qquad s.t. \quad \Gamma(M^H_{E_{k-1}}) = \Gamma(M^H) \cup E_{k-1}, \\ &\qquad\qquad E_{k-1} = \bigcup_{1}^{k-1} e_i, \text{ and } E_{k-1} \subseteq \Gamma(M^R) \end{aligned} \tag{2}$$

The search for OEG-NA naturally starts from $M^H_{E_{k-1}}$ for $\langle e_k, t_k \rangle$ since we no longer worry about matching the prefix.

C. OEG for Matching Any Prefix (OEG-AP)

One assumption made in both OEG-PP and OEG-NA is that the optimal plan for a given I, G pair is always unique. When we estimate the human's expected plan under a candidate model $M^H_{E_k}$ while searching for the kth sub-explanation, this assumption allows us to use the plan $\pi^H_{E_k}$ returned by any optimal planner, since they will always be the same. $\pi^H_{E_k}$ is then compared against $\pi^*_{I,G}$ to determine whether the e_k (incorporated into $M^H_{E_k}$) satisfies the requirements of OEG. When multiple optimal plans are present, the above check only needs to work for one of those plans. In this variation, we relax the uniqueness assumption of the optimal plans.

Definition 6 (OEG-AP): An online explanation for matching any prefix is a set of sub-explanations in the form of $\langle e_k, t_k \rangle$ such that the following is satisfied:

$$\begin{aligned} &\exists \pi^H_{E_{k-1}} \in \Pi^H_{E_{k-1}} \\ &\forall k > 0, \textit{Prefix}(\pi^*_{I,G}, t_k - 1) = \textit{Prefix}(\pi^H_{E_{k-1}}, t_k - 1) \\ &\qquad s.t. \quad \Gamma(M^H_{E_{k-1}}) = \Gamma(M^H) \cup E_{k-1}, \\ &\qquad\qquad E_{k-1} = \bigcup_{1}^{k-1} e_i, \text{ and } E_{k-1} \subseteq \Gamma(M^R) \end{aligned} \tag{3}$$

where $\Pi^H_{E_{k-1}}$ represents the set of all optimal plans under $M^H_{E_{k-1}}$. A similar definition can be provided for OEG-NA after removing the uniqueness assumption.

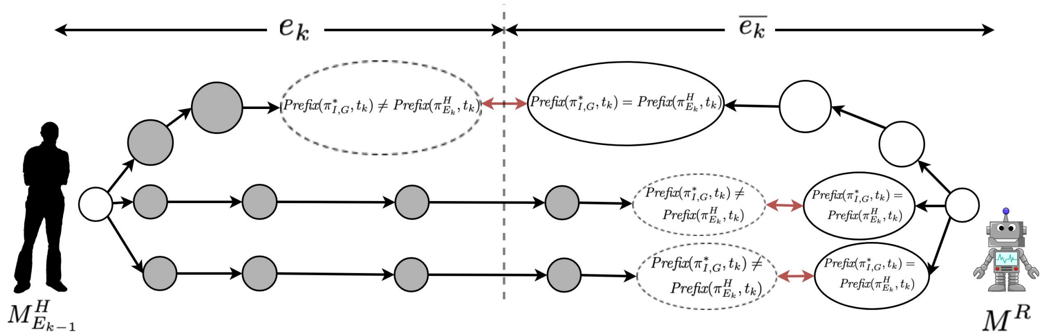

Fig. 2: The model space search process for the kth sub-explanation in OEG-PP. The search starts from M^R (similar to that used for MCE in [7]) until finding the largest set of $\overline{e_k}$ (or smallest e_k) that satisfies *Prefix*$(\pi^*_{I,G}, t_k)$ = *Prefix*$(\pi^H_{E_k}, t_k)$, under any M that is in between $M^R \setminus \overline{e_k}$ and M^R. Each node represents a candidate model and each edge a unit feature change. The gray nodes are nodes that are not expanded in the search.

To check for a candidate e_k, according to our previous discussion, we need to search for the largest set of model changes to M^R, denoted as $\overline{e_k}$, such that *Prefix*$(\pi^*_{I,G}, t_k)$ = *Prefix*$(\pi^H_{E_k}, t_k)$. An obvious solution to OEG-AP is to obtain $\Pi^H_{E_k}$ by computing all the optimal plans under $M^H_{E_k}$. This approach however is computationally expensive. Instead, we implement a compilation approach. In this approach, to check the above condition, we only need to solve two planning problems. The first planning problem is simple: finding an optimal plan $M^H_{E_k}$ under the given I, G. We denote the returned plan by any optimal planner as $\pi^H_{E_k}$ as usual. The second one is trickier in which we need to obtain a problem under $M^H_{E_k}$ such that *any* optimal plan would have to satisfy the condition *Prefix*$(\pi^*_{I,G}, t_k)$ = *Prefix*$(\pi^H_{E_k}, t_k)$. We denote the plan returned as $\hat{\pi}^H_{E_k}$. Now, we know that if the cost of $\hat{\pi}^H_{E_k}$ is equal to that of $\pi^H_{E_k}$, there must exist an optimal plan in the human's model that matches the prefix of the robot's plan. Otherwise, no such plan exists and a sub-explanation must be made. Hence, the key here is to ensure that a given plan prefix is always satisfied in a compiled model.

It turns out that this is not difficult to achieve. For all $a_i, a_{i+1} \in$ *Prefix*$(\pi^*_{I,G}, t_k)$, where a_i, a_{i+1} are two consecutive actions in $\pi^*_{I,G}$, the compilation can be achieved by adding a predicate p_i to a_i as an effect, which is also added as a precondition for a_{i+1}. a_{i+1}, in its turn, adds p_{i+1} as an effect which is a precondition for a_{i+2}, etc. The search process is the same as that described in OEG-PP. The search stops when any optimal plan in the human's updated model matches the robot's plan. In contrast to OEG-PP, the plan that is returned by an optimal planner under the human's model after an OEG-AP may not be exactly the robot's plan.

V. Evaluation

We evaluate our methods for online explanation generation both synthetically and in simulation with human subjects, and compare them with variations of minimally complete explanations (MCE) [7] as baselines. For the synthetic evaluation, our aim is to show how online explanations differ from MCEs. We evaluate our methods and MCE on 10 different problems across the IPC Rover and Barman domains [28]. For human subject study, our aim is to verify the following:

- *Online explanations reduce mental workload and improve task performance.*

A modified rover domain (Sec. V-B) is used. In all evaluations, M^R is the true domain model, and M^H is created by removing model features from M^R. All results are collected on a 2.2 GHz quad core Macbook Pro with 16 GB RAM.

A. Simulation Results

Table I presents the simulation results comparing OEG-PP, OEG-NA and OEG-AP with MCE. The benefits of online explanations are clear: the average size of sub-explanations is significantly smaller than the size of MCE, although the sum of their sizes is generally larger than the size of MCE. This shows that most explanations can indeed be broken up and communicated incrementally while subject to the requirements of online explanations! The effect of OEG-AP on the size of explanations is interesting, which suggests that removing the uniqueness assumption of the optimal plan has a positive impact on explanation generation: the sum of sub-explanations has a size that is smaller than MCE. This intuitively makes sense since not all the sub-explanations in MCE may be required as long as the robot's plan is optimal in the updated human's model (but differs from the plan found there by an optimal planner). To see the influence of removing the uniqueness assumption from another angle, for both OEG-NA and OEG-AP, we evaluate how the human's expected plan ($\pi^H_{E_k}$) after the explanation (returned by an optimal planner) may be different from $\pi^*_{I,G}$ using action plan distance, which has a value between 0 (no difference) and 1 (maximum difference). For OEG-NA, this distance is generally non-zero since only the very next action is considered when making a sub-explanation. For OEG-AP, the distance is also non-zero in general but due to the non-uniqueness of the optimal plan. Computationally, OEG methods are generally a bit faster than MCE which may appear to be surprising. Some analysis reveals that this is due to the fact that the incremental search in online explanation generation in fact reduces the search space by removing candidate features to be added to M_H for later searches. For OEG-NA and OEG-AP, this may also be due in part to the fact that they often terminate earlier and before $\pi^H_{E_k}$ becomes exactly $\pi^*_{I,G}$.

Pr.	OEG-PP		OEG-NA			OEG-AP			MCE	
	$\sum e_k/\lvert e_k\rvert$	Time	$\sum e_k/\lvert e_k\rvert$	Dist.	Time	$\sum e_k/\lvert e_k\rvert$	Dist.	Time	$\lvert E\rvert$	Time
Rover										
P1	3/1.5	8.9	7/1.2	0.40	17.9	2/1.0	0.40	6.9	3	28.9
P2	5/1.7	22.3	7/1.4	0.11	42.6	3/1.0	0.11	18.3	5	150.5
P3	6/1.5	18.7	8/1.1	0.07	21.3	3/1.0	0.07	1.6	5	176.2
P4	6/1.5	51.0	8/1.3	0.13	94.8	5/1.3	0.13	45.4	6	314.2
P5	5/1.7	54.8	8/1.3	0.14	106.7	3/1.5	0.14	50.4	4	272.8
Barman										
P1	5/1.3	43.0	5/1.3	0.91	59.9	2/1.0	0.94	24.4	5	180.0
P2	5/1.0	36.2	5/1.0	1.00	33.0	3/1.0	0.90	9.4	5	38.9
P3	5/1.3	36.8	5/1.0	0.90	46.8	3/1.5	0.71	9.7	5	51.8
P4	5/1.3	78.4	5/1.0	0.84	69.0	4/1.0	0.56	20.4	5	61.9
P5	5/1.7	41.9	5/1.0	0.89	54.7	3/1.5	0.56	10.2	5	61.5

TABLE I: Comparison of explanation size, average sub-explanation size (for online only), plan distance between $\pi^H_{E_k}$ and $\pi^*_{I,G}$ (when applicable) and time (in seconds) using the different methods for the IPC Rover and Barman domains.

B. Human Study

To test our hypothesis, we compare the explanations created by our methods with variations of MCE methods in a modified rover domain. The task is for the rover to collect and analyze soil and rock samples, take pictures of targets, and send them to the lander. To ensure that the performance difference is not solely due to breaking up the information, we implement another baseline that randomly breaks up an MCE into multiple parts and communicates each part separately so that they are uniformly distributed through the plan execution (referred to as MCE-R).

We conducted our experiment using Amazon Mechanical Turk (MTurk) with a 3D simulation of the rover domain (see Fig. 3). The subjects were first given an introduction to the rover domain and the task they were supposed to help with. In the experiment, we deliberately removed certain information from the introduction. In particular, we did not inform them that the storage space and memory of the rover is limited, the camera must be calibrated, and calibrated with respect to the target before taking an image. These introduced the differences between M^H and M^R.

Each subject was given a 30-minute limit to finish the task. Explanations were provided using plain English language and the rover actions were depicted using GIF images in the 3D simulation as the rover executed the plan. The human subject acted as the rover's supervisor, and was asked to determine whether each of the rover's action was questionable or not. Random actions were added into the plan to make sure that the subject must question some actions to perform well. Each subject was only allowed to perform the task for one setting (OEG-PP, OEG-NA, OEG-AP, MCE, or MCE-R) to reduce the influence of learning from repeated runs. To simulate highly demanding tasks, we have incorporated three spatial puzzles as secondary tasks. At the end of the study, the subjects were provided the NASA TLX to evaluate the workload [29] under several categories [30].

Results: We created the surveys using Qualtrics and recruited 150 human subjects on MTurk, with 30 subjects for each setting. To improve the quality of the responses, we set the criteria that the worker's HIT acceptance rate must be greater than 98%. After filtering out invalid responses (that failed to identify the 2 purposely inserted random actions out of a total of 30 actions in the plan), we obtained 94 valid responses in total: 19 for each of MCE-R and MCE, 20 for OEG-PP, and 18 for each of OEG-NA and OEG-AP. Their ages ranged from 18 to 70, and 29.8% of them were female.

Fig. 3: The 3D visualization of the modified rover domain. There are four robots on Mars, each has a different camera resolution and sampling equipment. The mission is to sample soil, rock and take images at different locations and communicate it to the lander shown on the right side of the picture.

	MCE-R	MCE	OEG-PP	OEG-NA	OEG-AP	Random
Accuracy	0.746	0.804	**0.858**	**0.852**	**0.872**	
# Actions	8.789	7.263	**5.250**	**5.330**	**4.940**	2.0/30

TABLE II: The accuracy and number of questionable actions based on the subjects' feedback for the five settings. The ground truth for questionable actions is 2 out of 30 in total.

The results show that OEGs in general performed significantly better than the baselines in both objective (Table II) and subjective measures (Fig. 4). Table II shows that the numbers of questionable actions are significantly lower for OEGs than MCEs (with p-values < 0.001). This indicates that the subjects had more trust towards robots in the OEG settings. The accuracy for identifying the correct actions (questionable vs. non-questionable) is also higher for OEGs (with p-values < 0.001). Among the three OEG methods, OEG-AP performed the best but no significant differences were observed in either the objective or subjective measures. This seems to suggest that the performances were dominated mainly by the average size of sub-explanations, which did not vary much among the OEGs (i.e., $\sum_{e_k}/|e_k|$: OEG-PP

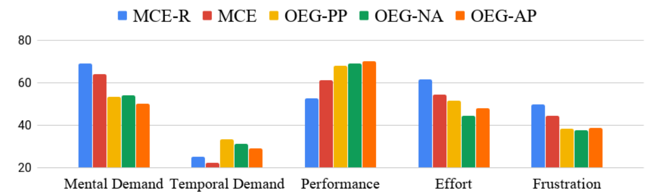

Fig. 4: Comparison of TLX categories for the five settings.

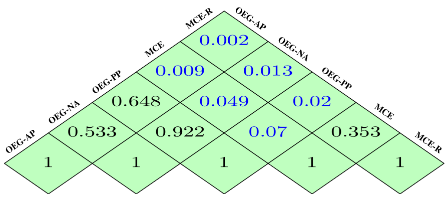

Fig. 5: p-values for the weighted sum of the subjective measures, with weights 1.0 for all TLX categories.

6/1.5, OEG-NA 5/1.25, OEG-AP 3/1.0, MCE 5/NA, MCE-R 5/1.0).

It is worth noting that MCE-R performed *worse* than MCE objectively with p-values 0.043 and 0.028 respectively for the two measures in Table II, which suggests that the performance difference was unlikely due to simply breaking up the information, thus confirming the usefulness of OEGs. The subjective measures in Fig. 4 for the most part reaffirm the conclusions. Due to intertwining explanations with plan execution, OEGs are expected to create more temporal demand. The p-values for the subjective measures are presented in Fig. 5. The results indicate statistically significant differences between OEGs and MCEs. The group-wise p-value is 0.0068 between OEGs and MCEs.

VI. Conclusions

In this paper, we introduced a novel formulation for explanation generation that was focused on reducing the mental workload for the human to interpret an explanations. We took a step further from prior work, which considered only the correct explanations, by proposing explanations that were also easily understandable. We provided three methods and evaluated them both in simulation and with human subjects. Results confirmed that they improved task performance and reduced mental workload.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their helpful comments. This research is supported in part by the NSF grant IIS-1844524, the NASA grant NNX17AD06G, and the AFOSR grant FA9550-18-1-0067

References

[1] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.
[2] M. R. Endsley, "Design and evaluation for situation awareness enhancement," in *Proceedings of the Human Factors Society annual meeting*, vol. 32. SAGE Publications Sage CA: Los Angeles, CA, 1988, pp. 97–101.
[3] N. J. Cooke, "Team cognition as interaction," *Current directions in psychological science*, vol. 24, no. 6, pp. 415–419, 2015.
[4] M. Göbelbecker, T. Keller, P. Eyerich, M. Brenner, and B. Nebel, "Coming up with good excuses: What to do when no plan can be found," in *ICAPS*, 2010.
[5] M. Hanheide, M. Göbelbecker, G. S. Horn, A. Pronobis, K. Sjöö, A. Aydemir, P. Jensfelt, C. Gretton, R. Dearden, M. Janicek *et al.*, "Robot task planning and explanation in open and uncertain worlds," *Artificial Intelligence*, vol. 247, pp. 119–150, 2017.
[6] S. Sohrabi, J. A. Baier, and S. A. McIlraith, "Preferred explanations: Theory and generation via planning," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.
[7] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," *IJCAI*, 2017.
[8] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.
[9] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, "Plan explicability and predictability for robot task planning," in *ICRA*. IEEE, 2017, pp. 1313–1320.
[10] B. B. Fischer and L. Fischer, "Styles in teaching and learning," *Educational leadership*, vol. 36, no. 4, pp. 245–254, 1979.
[11] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.
[12] N. Cowan, "What are the differences between long-term, short-term, and working memory?" *Progress in brain research*, vol. 169, pp. 323–338, 2008.
[13] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.
[14] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Twenty-Ninth IAAI Conference*, 2017.
[15] G. Brewka, "Artificial intelligence—a modern approach by stuart russell and peter norvig, prentice hall. series in artificial intelligence, englewood cliffs, nj." *The Knowledge Engineering Review*, vol. 11, no. 1, pp. 78–79, 1996.
[16] R. S. Andersen, O. Madsen, T. B. Moeslund, and H. B. Amor, "Projecting robot intentions into human environments," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 294–301.
[17] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 301–308.
[18] Z. Gong and Y. Zhang, "What is it you really want of me? generalized reward learning with biased beliefs about domain dynamics." in *AAAI*, 2020, pp. 2485–2492.
[19] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang, "Ai challenges in human-robot cognitive teaming," *arXiv preprint arXiv:1707.04775*, 2017.
[20] M. Zakershahrak, A. Sonawane, Z. Gong, and Y. Zhang, "Interactive plan explicability in human-robot teaming," in *RO-MAN*. IEEE, 2018, pp. 1012–1017.
[21] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," *arXiv preprint arXiv:1709.10256*, 2017.
[22] S. Reddy, A. Dragan, and S. Levine, "Where do you think you're going?: Inferring beliefs about dynamics from behavior," in *NeurIPS*, 2018, pp. 1461–1472.
[23] Z. Gong and Y. Zhang, "Behavior explanation as intention signaling in human-robot teaming," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1005–1011.
[24] Y. Zhang and M. Zakershahrak, "Progressive explanation generation for human-robot teaming," *arXiv preprint arXiv:1902.00604*, 2019.
[25] M. Fox and D. Long, "Pddl2. 1: An extension to pddl for expressing temporal planning domains," *Journal of artificial intelligence research*, vol. 20, pp. 61–124, 2003.
[26] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3-4, pp. 189–208, 1971.
[27] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational psychologist*, vol. 38, no. 1, pp. 1–4, 2003.
[28] IPC, "Competition domains for international planning committee," http://www.plg.inf.uc3m.es/ipc2011-learning/Domains.html, Jul. 2019.
[29] NASA, "Nasa task load index," https://humansystems.arc.nasa.gov/groups/TLX/, Feb. 2019.
[30] P. S. Tsang and V. L. Velazquez, "Diagnosticity and multidimensional subjective workload ratings," *Ergonomics*, vol. 39, no. 3, pp. 358–381, 1996.