Online Explanation Generation for Human-Robot Teaming

Mehrdad Zakershahrak, Ze Gong, Akkamahadevi Hanni and Yu Zhang
School of Computing, Informatics and Decision Systems Engineering
Arizona State University, Tempe, AZ 85281 USA
{ mzakersh,zgong11,ahanni,Yu.Zhang.442}@asu.edu

Abstract

As Artificial Intelligence (AI) becomes an integral part of our life, 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 explain its behavior is one of the key requirements of an explainable agency. Prior work on explanation generation focuses on supporting the reasoning behind the robot's behavior. These approaches, however, fail to consider the cognitive effort needed to understand the received explanation. In particular, the human teammate is expected to understand any explanation provided before the task execution, no matter how much information is presented in the explanation. In this work, we argue that an explanation, especially complex ones, should be made in an online fashion during the execution, which helps to spread out the information to be explained and thus reducing the cognitive load of humans. However, a challenge here is that the different parts of an explanation are 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. We base our explanation generation method in a model reconciliation setting introduced in our prior work. Our approach is evaluated both with human subjects in a standard planning competition (IPC) domain, using NASA Task Load Index, as well as in simulation with ten different problems.

Introduction

One of the most important aspects of human-AI interaction is for the AI agent to provide explanations to convey the reasoning behind the robot's decision-making (Lombrozo 2006). An explanation provides insights about the agent's intent, which helps the human maintain trust of the robotic peer as well as a shared situation awareness (Endsley 1988; Cooke 2015). Prior work on explanation generation often focuses on supporting the motivation for the agent's decision while ignoring the underlying requirements of the recipient to understand the explanation (Göbelbecker et al. 2010; Hanheide et al. 2017; Sohrabi, Baier, and McIlraith 2011). A good explanation should be generated in a lucid fashion from the recipient's perspective (Chakraborti et al. 2017).

In our prior work (Chakraborti et al. 2017), we encapsulate such inconsistencies as *model differences*, while considering the discrepancies between the human and its own model when generating explanations. An explanation then becomes a request to the human to adjust the model differences in his mind so that the robot's behavior would make sense in the updated model, which captures the human's expectation of the robot. The general decision-making process of an agent in the presence of such model differences is termed *model reconciliation* (Chakraborti et al. 2017; Zhang et al. 2017).

One remaining issue, however, is the ignorance of the cognitive effort required of the human for understanding an explanation. In previous work, the human is expected to understand any explanation provided before the task execution, regardless of how much information is present. In this work, we argue that explanations, especially complex ones, should be provided in an online fashion, which intertwines the communication of explanations with plan execution. In such a manner, an online explanation requires less cognitive effort at any specific point of time. The challenge here, however, is that the different parts of an explanation are dependent on each other, which must be taken into account when generating online explanations. The online explanation generation process spreads out the information to be communicated while ensuring that they do not introduce cognitive gaps so that the different parts of the information are perceived in a smooth fashion.

Motivating Example: Let us illustrate the concept of online explanations, by considering a familiar situation where two friends, Mark and Emma, want to meet up to study together for an upcoming exam. Mark is a take-it-easy person so he plans to break the review session into two 60 minutes parts, grab lunch in between the sub-sessions and go for a walk after lunch. On the other hand, Mark knows that Emma is of a focused type who would rather keep the review in one session and get lunch afterwards. Mark would like to keep his plan. However, had he explained to Emma at the beginning of his plan, he knew that Emma would have proposed to order takeout for lunch on the way before the review session. Instead, without revealing his plan, he goes with Emma to the library. After studying for 60 minutes, he then explains to Emma that he cannot continue without energy, which makes going to lunch the best option for both. However, Mark refrained from telling Emma that he also needed a walk (until after lunch) since otherwise Emma would have proposed for him to take a walk alone while she stays a bit longer for

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

review, and meet up at the lunch place.

The above example demonstrates the importance of providing parts of an explanation in an online fashion. Mark gradually reveals the reasoning to maintain his plan as the execution unfolds so that it also becomes acceptable (and understandable) to Emma. The key point here is to explain minimally and only when necessary, as long as the next action becomes understandable. In this way, the information to be conveyed is spread out into the future so that there is less cognitive requirement at the current step–from Emma's perspective, the interaction with Mark is simpler and requires less thought.

In this paper, we develop a new method for explanation generation that intertwines explanation with plan execution. The new form of explanation is referred to as *online explanation*, which considers the cognitive effort of the receiver of an explanation by breaking it into multiple parts that are to be communicated at different times during the plan execution. Each part of an explanation is responsible for clarifying a prefix of a robot's plan to be explained, so that the prefix makes sense to the receiver. The entire robot's plan will only become explainable when all parts of an explanation are provided. We use a model search method that ensures that the earlier parts communicated would not affect the later parts of the explanation. This creates a desirable experience for the receiver of an explanation by significantly reducing the cognitive effort. Our approach is evaluated both with human subjects and in simulation.

Related Work

Recently, explainable AI paradigm (Gunning 2017) rises as one essential constituent of human-AI collaboration. The explainable agent's effectiveness (Langley et al. 2017) is assessed based on its capability to model the human's perception of the AI agent accurately. This model of the other agents allows the agent to infer about their expectation of itself. Using this model, an agent can generate legible motions (Dragan, Lee, and Srinivasa 2013), explicable plans (Zhang et al. 2017; Zakershahrak et al. 2018; Fox, Long, and Magazzeni 2017), signal its intention before execution (Gong and Zhang 2018) or assistive actions (Reddy, Dragan, and Levine 2018). In these approaches, an agent often substitutes cost optimality with a new metric that simultaneously considers cost and explicability.

Another approach of using this model is for the agent to explain its behavior by generating explanations, while considering the agent can maintain its optimal behavior (Göbelbecker et al. 2010; Hanheide et al. 2017; Sohrabi, Baier, and McIlraith 2011). All of these approaches assume that the provided explanation would be understood, regardless of the timing and/or the amount of information being shared–the cognitive effort that is required for understanding an explanation is largely ignored. In our prior work, we have studied how the ordering of the information of an explanation may influence the perception of an explanation (Zhang and Zakershahrak 2019). In this work, we further argue that an explanation, especially complex explanations that require a large amount of information to be conveyed, must sometimes be made in an online fashion. The idea behind online explanation generation thus is to provide a minimal amount of information that is sufficient to explain the next action,

Explanation Generation

Since our problem definition is based on the model reconciliation setting defined in our prior work (Chakraborti et al. 2017), we briefly review relevant concepts before defining our problem in this work. Our problem is closely associated with planning problems, which are defined as a tuple $(F, A, \mathcal{I}, \mathcal{G})$ using PDDL (Fox and Long 2003), similar to STRIPS (Fikes and Nilsson 1971). F is the set of predicates used to specify the state of the world and A is the set of actions used to change the state of the world. Actions are defined with a set of preconditions, add and delete effects. $\mathcal{I}, \mathcal{G}$ are the initial and goal state.

Definition 1 (Model Reconciliation) *: A model reconciliation (Chakraborti et al. 2017) is a tuple* $(\pi^*_{I,G}, \langle M^R, M^H \rangle)$*, where* $cost(\pi^*_{I,G}, M^R) = cost^*_{M^R}(I, G)$ *and* $\pi^*_{I,G}$ *is the robot's plan to be explained.*

$cost(\pi^*_{I,G}, M^R)$ is the cost of the plan generated using M^R and $cost^*_{M^R}(I, G)$ is the cost of the optimal plan based on the initial and goal state pair under M^R. Hence, the robot plan to be explained is required to be optimal according to M^R, assuming rational agents. When the robot's behavior to be explained (i.e., $\pi^*_{I,G}$) matches with the human's expected behavior, the models are said to be reconciled for the plan. Using M^H, the human generates π_{M^H}, which captures the human's expectation of the robot. Whenever the two plans are different, the robot should explain by generating an explanation to reconcile the two models. Explanation generation in a model reconciliation setting means bringing two models (M^H and M^R) *"close enough"* such that $\pi^*_{I,G}$, the robot's plan, becomes fully explainable (optimal) in the human's model.

In (Chakraborti et al. 2017), the following mapping function over a planning problem is defined to convert a planning problem into a set of features that specifies the problem as Γ: $\mathcal{M} \longmapsto$ S' is a mapping function, which transfers any planning problem $(F, A, \mathcal{I}, \mathcal{G})$ to a state s' in the feature space as follows:

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

$$\Gamma(\mathcal{M}) = \{\tau(f) | \forall f \in \mathcal{I} \cup \mathcal{G} \cup \bigcup_{a \in A} \{f' | \forall f' \in \{c_a\} \cup pre(a) \cup eff^+(a) \cup eff^-(a)\}\}$$

Definition 2 (Explanation Generation) *The explanation generation problem (Chakraborti et al. 2017) is a tuple* $(\pi^*_{I,G}, \langle M^R, M^H \rangle)$*, and an explanation is a set of unit feature changes to* M^H *such that* 1) $\Gamma(\widehat{M^H}) \setminus \Gamma(M^H) \subseteq \Gamma(M^R)$*, 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 two models by making the cost difference between the human's expected plan and

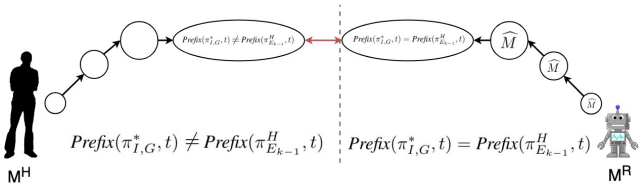

Figure 1: Model space search process for OEG.

the robot's plan smaller after the model updates.

Definition 3 (Complete Explanation) *Given an explanation generation problem, a complete explanation (Chakraborti et al. 2017) is an explanation that satisfies* $cost(\pi^*_{I,G}, \widehat{M^H}) = cost^*_{\widehat{M^H}}(I, G)$.

In other words, the robot's plan must be optimal in the human's model after a complete explanation ($\widehat{M^H}$). A minimal complete explanation (MCE) is defined as a complete explanation that contains the minimum number of unit feature changes.

Online Explanation Generation (OEG)

We introduce online explanation to consider the cognitive requirement of the human for understanding the generated explanation by the agent. The key here is to only provide a minimal amount of information to explain the next action that is not explainable.

Problem Formulation

Definition 4 (Online Explanation Generation) *An online explanation is a set of sub-explanations* (e_k, t_k)*, where* e_k *represents the* k*th set of unit features to be made (as a sub-explanation) at step* t_k *in the plan, such that:*

$$\forall k > 1, \forall t < t_k, Prefix(\pi^*_{I,G}, t) = Prefix(\pi^H_{E_{k-1}}, t)$$
$$a_t \in \pi^*_{I,G} \ \& \ M^H_{E_{k-1}} = \Gamma(M^H \wedge E_{k-1}) \quad (1)$$

where $Prefix(\pi, t)$ returns the prefix of a plan π up to step t. E_k represents $e_{1:k}$ and $\pi^H_{E_k}$ is the plan created from $M^H_{E_k}$ (M^H after providing sub-explanations e_1 to e_k).

Basically, an online explanation requires only that any actions in the robot's plan before the kth sub-explanations will match with that of the human's expectation. In such a way, the robot can split an explanation into multiple parts, which are made in an online fashion as the plan is being executed.

OEG with Model Space Search

We convert the problem of explanation generation to the problem of model search in the space of possible models. To generate the sub-explanations (i.e., $\{e_k\}$) for an online explanation, similar to the search process for complete explanations (Chakraborti et al. 2017), we convert the problem of explanation generation to the problem of model search in the space of possible models. Therefore, the planning process must consider how the sequence of model changes would result in the changes of the human's expectations after each sub-explanation. The challenge here is that the model changes may not be independent, i.e., future changes may render a mismatch in the previously reconciled plan prefixes. To address this issue, it must be ensured that the model changes after e_k, i.e., $e_{k+1:m}$ where m denotes the size of the set of sub-explanations, would not change the plan prefixes in M^H. This can be achieved by searching from M^R to M^H to find the largest set of model changes which ensure the plan prefix would not change afterwards after further sub-explanations. This search process is illustrated in Figure 1.

More specifically, the following process will be performed recursively for each sub-explanation. First, we continue moving along $\pi^*_{I,G} = (a_1, a_2, ..., a_n)$ as long as the plan prefix matches with the prefix of the plan using the human model M^H. Let $t = t_1$ be the first plan step where they differ. Our search for the sub-explanation starts with M^R. It finds the largest set of model changes (λ) to M^R such that the prefix of a plan using the corresponding model (i.e., M^R minus the set of changes) matches with that of $\pi^*_{I,G}$ up to step t_1. The complement set of changes (i.e., the difference between M^H and M^R minus λ) will be e_1. For the next recursive step, we will start from action $t_1 + 1$ and the human model will be $M^H_{E_1}$. To ensure that the prefix (up to t_1) will be maintained for future steps, we directly force the later plans to be compatible with the prefix. Since we know that an optimal plan exists that satisfies this requirement following the search process, this would not affect our solution for online explanation.

Algorithm 1: OEG-Search for e_k

input : $M^H_{E_{k-1}}$, M^R, $\pi^*_{I,G}$ and $\{e_{k-1}, t_{k-1}\}$
output: Sub-explanation e_k
Compute $\Delta(M^H_{E_{k-1}}, M^R)$ as the difference between the two models;
Compute π_H based on $M^H_{E_{k-1}}$ with prefix set up to t_{k-1};
$t_k \leftarrow$ FirstDiff($\pi^*_{I,G}$, π_H);
LONGESTMONOTONIC($M^H_{E_{k-1}}, k$)
if ($\pi^*_{I,G} \equiv \pi_H$) **then**
  return $\{\}$;
for $\forall f \in \Gamma(M^R) \backslash \Gamma(M^H_{E_{k-1}})$ **do**
  ▷ All remaining differences after sub-explanations E_{k-1}
  $\lambda \leftarrow \Gamma(\overline{M^H_f})$; ▷ create a modification
  Create a plan π^f_H using $(\overline{M^H_f})$;
  if ($t_k \leq$ *FirstDiff*(π^f_H, $\pi^*_{I,G}$)) **then**
    if $|\lambda| > \lambda_{max}$ **then**
      $\lambda_{max} \leftarrow \lambda$
return λ_{max} as e_k

Algorithm

The search algorithm for OEG is presented in Algorithm 1 for finding e_k given E_{k-1}. To search for e_k, we use a breadth-first search procedure on the model space. Given $M^H_{E_{k-1}}$ and M^R, we start off with finding the difference between these two models, and modify M^R with respect to

M^H to find the largest set of model changes that can satisfy constraint introduced in Eq. (1). This algorithm continues until the human's plan matches with that of the robot's plan.

Evaluation

We evaluated our approach for online explanation generation with human subjects and in simulation. For simulation, the goal is to see that online explanation is in general different from minimally complete explanation(MCE) in terms of the information needed. We evaluated our approach on ten different problems in the rover domain–a standard IPC domain described below. For human subject study, the aim here is to confirm the benefits of online explanation generation. Our hypothesis is that online explanation generation will reduce cognitive load and improve task performance. We evaluated our approach with human subjects on a slightly modified rover domain that is more complex.

Rover Domain

In this domain, the rover is supposedly on Mars and the goal is to explore the space to take rock and soil samples as well as taking images and communicate the results after analysis to the base station using the lander. In order to take any image, the rover must first calibrate its camera with respect to the target objective. To sample rock or soil, the robot must have an empty space in its storage. At any point of time, the rover only has enough space to store one sample. In order to take multiple samples, it must drop the current sample (after analysis) before taking another sample. For the evaluation with human subject, we further restrict the memory size of the rover for storing images so that if a new image is taken, an old image will be overwritten, so the rover needs to transfer an image before taking another one.

Simulation Results Table 1 shows the simulation results comparing MCE versus our approach within 10 problems in rover domain. While the total number of predicates that are included in our approach is more than that in MCE, the average sub-explanation size in our approach is 1, meaning that the number of predicates for all sub-explanations are 1. This is expected to have a positive effect on the human's cognitive load, which we evaluated next.

Human Study

To test our hypothesis, we compared our approach for OEG with MCE (Chakraborti et al. 2017) using the rover domain. We conducted our experiment using Amazon Mechanical Turk (MTurk). The subjects were given an introduction to the rover domain . Then they were asked to help the rover create a plan for the rover with OEG or MCE. Each subject can only perform the task for one setting to reduce the influence between different runs. Moreover, we deliberately remove certain information about the domain so that the subject would create a wrong plan when no explanation is given. This hidden information introduces differences between M^H and M^R in the model reconciliation setting, and hence resulting in scenarios where explanations must be provided. This information, for example, includes the calibration of camera to take an image of every subject they were told to. Figure 2 presents the scenario of the rover problem used in our study.

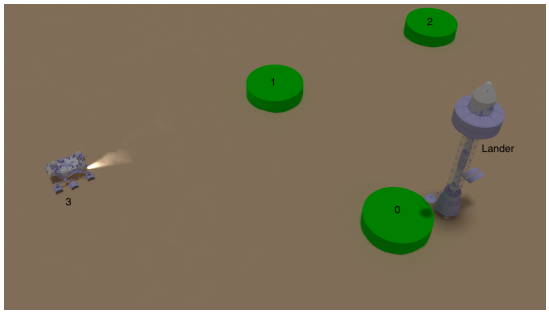

Figure 2: An illustration of rover problem provided to human subjects.

Problem #	1	2	3	4	5	6	7	8	9	10
MCE	3	2	5	5	6	5	6	5	4	4
OEG	4	3	6	7	8	8	7	6	6	7

Table 1: Simulation Results for IPC Rover Domain

At the end of the study, the subjects were provided the NASA Task Load standard questionnaire to evaluate the efficiency of our approach by NASA Task Load Index (TLX) (NASA 2019) which is a subjective workload assessment tool to evaluate human-machine interface systems. It calculates an overall workload score using a weighted average on sub-scales: mental demand, physical demand, temporal demand, performance, effort and frustration. The questions for each category and their weight assigned to each category is described below:

- How mentally demanding was the task? (5)
- How hurried or rushed was the pace of the task? (4)
- How successful were you in accomplishing what you were asked to do? (3)
- How hard did you have to work to accomplish your level of performance? (2)
- How insecure, discouraged, irritated, stressed, and annoyed were you? (1)

We used 100-point raw scales with the minimum gradient of 5 for each category for this questionnaire.

Results We recruited 30 human subjects on MTurk for each setting. After removing invalid plans that are too short (i.e., less than 4 plan steps), we obtained 14 responses for the MCE setting and 11 responses for OEG. The comparison of the TLX results of two settings is presented in Figure 3. The results show that OEG is less mentally demanding than complete explanations which suggests that OEG can indeed reduce the human's mental load effectively. Furthermore, the human subjects' also reported that they were more comfortable with their performances with online explanation. In terms of frustration level of the human subjects, OEG seems to have discouraged them less than MCE as well.

Further, we calculate the adjusted rating by multiplying the weight to the rate given by the human subjects. OEG has a weighted rating of 48.72 while the MCE is 49.36. A larger weighted rating indicates more load introduced to the human. The result shows that OEG is able to better reduce the human's load than MCE. Figure 3 presents the subjective results of our human study, across 5 TLX categories.

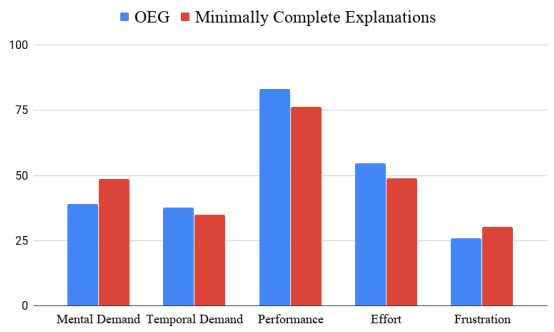

Figure 3: Comparison of TLX results for two settings.

Approach	Cronbach α	Std. α	G6 (smc)	Avg. R
OEG	0.9605	0.9605	0.9226	0.9605
MCE	0.9226	0.9605	NA	NA

Table 2: Cronbach Alpha Value Comparing Minimally Complete Explanations(MCE) and Online Explanations(OEG)

for both settings. We performed unpaired t-test over the two categories. Based on our results, $p = 0.2116$ for the mental demand. [1]

Lastly, we examine how well the human subjects understand the robot's plan given the different explanations. We compare the distances between the human's plan and the robot's plan by computing the ratio between the number of the actions in the human's plan that are also in the robot's plan and the number of actions in the robot's plan for both the OEG and MCE settings. The average distance value of OEG is 0.783 which outperforms 0.753 for the MCE. Since all of the five categories in our subjective study are inextricably related to the cognitive load, we calculated the Cronbach's Alpha values for all metrics in NASA TLX (Cronbach 1951) in Table 2.

Conclusion

In this paper, we introduced a novel approach for explanation generation to reduce the cognitive effort needed for the human to interpret the explanations, throughout a human-robot interaction scheme. The key idea here is to break down a complex explanation into smaller parts and convey them in an online fashion, while intertwined with the plan execution. We take a step further from our prior work by considering not only providing the correct explanations, but also the explanations that are easily understandable. This is important for achieving explainable AI. We evaluated our approach using both simulation and human subjects. Results showed that our approach achieved better task performance while reducing the cognitive load.

References

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*.

Cooke, N. J. 2015. Team cognition as interaction. *Current directions in psychological science* 24(6):415–419.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16(3):297–334.

Dragan, A. D.; Lee, K. C.; and Srinivasa, S. S. 2013. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, 301–308. IEEE Press.

Endsley, M. R. 1988. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting*, volume 32, 97–101. SAGE Publications Sage CA: Los Angeles, CA.

Fikes, R. E., and Nilsson, N. J. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence* 2(3-4):189–208.

Fox, M., and Long, D. 2003. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research* 20:61–124.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable planning. *arXiv preprint arXiv:1709.10256*.

Göbelbecker, M.; Keller, T.; Eyerich, P.; Brenner, M.; and Nebel, B. 2010. Coming up with good excuses: What to do when no plan can be found. In *ICAPS*.

Gong, Z., and Zhang, Y. 2018. Robot signaling its intentions in human-robot teaming. In *HRI Workshop on Explainable Robotic Systems*.

Gunning, D. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.

Hanheide, M.; Göbelbecker, M.; Horn, G. S.; Pronobis, A.; Sjöö, K.; Aydemir, A.; Jensfelt, P.; Gretton, C.; Dearden, R.; Janicek, M.; et al. 2017. Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence* 247:119–150.

Langley, P.; Meadows, B.; Sridharan, M.; and Choi, D. 2017. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*.

Lombrozo, T. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10(10):464–470.

NASA. 2019. Nasa task load index. https://humansystems.arc.nasa.gov/groups/TLX/.

Reddy, S.; Dragan, A.; and Levine, S. 2018. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In *NeurIPS*, 1461–1472.

Sohrabi, S.; Baier, J. A.; and McIlraith, S. A. 2011. Preferred explanations: Theory and generation via planning. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Zakershahrak, M.; Sonawane, A.; Gong, Z.; and Zhang, Y. 2018. Interactive plan explicability in human-robot teaming. In *RO-MAN*, 1012–1017. IEEE.

Zhang, Y., and Zakershahrak, M. 2019. Progressive explanation generation for human-robot teaming. *arXiv preprint arXiv:1902.00604*.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan explicability and predictability for robot task planning. In *ICRA*, 1313–1320. IEEE.

[1] Although we did not observe a significant difference here, we expect it to be due to insufficient data points collected.