Explain by Goal Augmentation: Explanation Generation as Inverse Planning

Zhiang Chen and **Yu Zhang**
Arizona State University
zch256@asu.edu, yzhan44@asu.edu

***Abstract*—There is a growing interest in achieving explainable robotics during human-robot interaction. However, existing work either focuses on modulating robot plans to make them more explicable, or uses explanations at the cost of increased human cognitive load. Prior work on generating explanations often requires updates to human domain knowledge. However, such work idealized both human memory capacity for updating knowledge and cognitive capacity for reasoning with the new knowledge. Inspired by psychological studies of human cognitive load, we aim at not changing human domain knowledge, which represents long-term memory that is difficult to update, and instead consider explanation generation as an inverse planning problem to search for a new goal that covers the original goal while resulting in an explicable plan. The new goal is then used as an explanation to the human observer. While our method for explanation generation is motivated by minimizing human cognitive load, it may also be used to hide private information from other agents, which can be useful in non-collaborative or adversarial settings. We introduce the problem formulation and illustrate the idea with a few examples.**

I. Introduction

As AI becomes ubiquitous, there is a growing interest in achieving explainable robotics during human-robot interaction. An important research topic along this direction is to enable the robot to demonstrate explainable behaviors, which often requires the robot to deviate from the optimal plans, and instead chooses plans that are more explicable to the human. When such plans are too costly to the robot, the robot is expected to explain its behavior. Prior work [1] formulates explanation generation as updates to the human's mental model of the robot D^H (i.e., an approximation of the robot domain model D^R), so that the robot's behavior becomes explicable to the human in this updated model. In particular, the problem of explanation generation is formulated as a model reconciliation problem, which aims at making changes to the human's mental model of the robot to bring it closer to the robot's domain model. Sreedharan et al. [2] extends the formulation to a hierarchical setting, and considers explanations in the framework of counterfactual reasoning, where propositions are searched to explain a foil that leads to a failure in achieving the goal.

Although these prior methods of explanation generation for the purpose of updating the human's mental domain D^H will theoretically work, one fatal issue is that the effectiveness of requiring updates to human knowledge is unpredictable. According to psychologic studies of human cognitive load [3], humans use both short-term and long-term memories, with the short-term memory being more volatile and changeable. However, the human's knowledge about the environment including other agents, such as our biases, falls more into the long-term memory, which is more reluctant to change. It is easy to change the human's mental model of the robot in an algorithm after making an explanation, but difficult in practice to ensure that the update actually happens in one's mind. This intuition is also backed up by psychological studies. As shown in [3] for example, when an adult is interacting with kids, even though the adult can provide sound explanations to warn the kids of dangerous situations, the kids can still make "mistakes" due to the limited abilities to update their knowledge, when their short-term memory limitation is reached.

A related issue that also affects the effectiveness of updating D^H is due to the limitation of the human's computational ability to reason with new knowledge. In contrast to a computer program, new knowledge cannot be simply plugged in and played with a human. It requires a long time to sink into one's mind, which involves a complex cognitive process that analyzes, dissects, and merges it with exiting knowledge. In the psychology literature, this computational limitation is studied in the ability of cognitive learning [4]. In human-robot teaming, more specifically, when a robot provides an explanation to the human, due to the limitation of human computational ability, the human may not be able to relate the explanation to the robot's behavior as the robot expects. That is, even though the explanation can be forced into the human's mind, the robot's behavior remains inexplicable although it would be when assuming unlimited computational ability.

Just to conclude, providing explanations to update human's knowledge relies on two unrealistic assumptions about humans: 1) the ability to update their knowledge without considering the short-term memory capacity and complexity of explanations; 2) the ability to reason with the newly updated knowledge base. In this work, we propose a new approach to explanation generation that does not involve changing the human's long-term knowledge while minimizing changes to the short-term memory, so that the problems aforementioned can be alleviated or often avoided. In particular, we consider explanation generation as an inverse planning problem to search for a new goal that also covers the original goal while also resulting in an explicable plan.

Next, we first provide more details about the related work and formulate our problem. We then illustrate explanation generation with a few examples and conclude afterwards.

II. Related work

Given the increased interest in achieving explainable robotics, research has been conducted along the direction in many areas, such as communication, motion planning and high-level task planning. In particular, AI challenges in human-robot cognitive teaming [5] have been surveyed from a task planning perspective. Several interesting approaches have been discussed there. For example, Zhang et al. [6] introduced the notion of plan explicability and built a learning model to compute it, which was then used by agents to generate plans that are more explicable to humans. This approach uses a similar principle as in this work, with the goal of reducing human cognitive load. However, this approach introduces additional constraints to planning and hence may not always be feasible (since the plan may be too costly). We show that our approach in fact generalizes plan explicability with slight modifications to the formulation. When generating explicable plans is infeasible, plan explanations as udpates to the human's knowledge can be considered as a model reconciliation problem [1] [2]. As discussed, however, these methods require strong assumptions that are unrealistic in practice.

Instead of updating the human's knowledge, our proposed method for generating explanations updates the robot's goal, which can be seen as an inverse planning problem, such that the original goal is covered by the new goal while resulting in an explicable plan to human. While in planning, we compute a plan to achieve a goal; in inverse planning, in contrast, we search for a goal that maximizes the explicability of the resulting plan. Our work also differs from plan recognition since no observations are required.

III. Methodology

Our problem setting is the one where the robot is cohabiting with humans and the robot has the ability to choose its own goals (of course, with the ultimate goal to make human lives betters, hopefully).

A. Background

A STRIPS planning problem is a tuple $M = \langle D, I, G\rangle$, where $D = \langle F, A\rangle$ is the domain model, F is the set of predicates, and $A = \{a\}$ is a set of actions with preconditions, add and delete effects: $Pre(a) \subseteq F$, $Add(a) \subseteq F$, and $Del(a) \subseteq F$. $I \subseteq F$ is the initial state, and $G \subseteq F$ is the goal state. A plan is a sequence of actions $\pi = (a_1, a_2, ..., a_n)$, where $\forall_i a_i \in A$. The cost of a plan π is given by $C(\pi) = \sum_{a\in\pi} c(a)$, where $c(a)$ is often a non-negative value; $C(\pi) = \infty$ if $\pi = \emptyset$ (no plan exists). A solution to a planning problem is a plan such that $\delta_D(I, \pi) \vDash G$, where $\delta_D(\cdot)$ is a transition function that transitions the state I using π. An optimal solution set of a problem is denoted as $\Pi^*_M = \{\pi^*_M | \pi^*_M = \text{argmin}_\pi[C(\pi)\, s.t.\, \delta_D(I, \pi) \vDash G]\}$. A plan in the optimal solution set is called an optimal plan $\pi^*_M \in \Pi^*_M$.

In human-robot teaming, we define a general explanation generation setting by the tuple $\langle M^R, M^H\rangle$, where $M^R = \langle D^R, I^R, G^R\rangle$ is the robot's model, and $M^H = \langle D^H, I^H, G^H\rangle$ is the human's mental model of the robot [1]. For simplicity, $\Pi^*_{M^R}$, $\Pi^*_{M^H}$ are denoted as Π^*_R, Π^*_H, and the same applies to π^*_R, π^*_H for $\pi^*_{M^R}$, $\pi^*_{M^H}$. Note that changes to all components of M is considered equivalent in [1].

B. Problem Formulation

We first provide a general definition of explicability used in our work, which is similar to that in [6]. Our approach does not depend on how explicability is defined.

Definition 1. *(Explicability) Given two planning problems with the same initial state and goal state, $M^R = \langle D^R, I, G\rangle$ and $M^H = \langle D^H, I, G\rangle$, and a plan π_R of M^R, the explicability of π_R to the model M^H is defined as $\xi_{[\pi_R:M^H]} = -min_{\pi^*_H}[dist(\pi_R, \pi^*_H)]$, where $dist(\cdot)$, similar to that in [6], computes the explicability distance between the two plans.*

We refer to this definition as the **Plan-to-Model Explicablility** (PME). Note that there are a couple of properties of the definition:

- PME is non-positive, $\xi_{max} = 0$
- $\xi_{[\pi^*_R:M^H]} = 0$ if $\Pi^*_R \cap \Pi^*_H \neq \emptyset$

In plan explanations as model reconciliation [1] [2], an explanation $\mathcal{E}$ is generated to update M^H to be $M^{H'}$ such that 1) $M^{H'}$ is closer to M^R, and 2) $\exists \pi^*_{H'} \in \Pi^*_R$.

Instead of changing everything in M^H, we aim at generating explanations as updating the goal G such that the new goal $G' = \text{argmax}_{G'}[\xi_{[\pi_{R'}:M^{H'}]}]\ s.t.\ G \setminus G' = \emptyset$
where $M^{H'} = \langle D^H, I, G'\rangle$ and $\pi_{R'} \in \Pi_{R'}$ is a solution for $M^{R'} = \langle D^R, I, G'\rangle$. To understand this, we firstly take a look at the optimization without the constraint. We expect that a plan $\pi_{R'}$ for the updated robot problem $M^{R'}$ to be completely explicable in the updated human model $M^{H'}$, that is, $\xi_{[\pi_{R'}:M^{H'}]} = 0$; if such a plan cannot be found, we search for a plan that can maximize $\xi_{[\pi_{R'}:M^{H'}]}$. At the same time, the new goal should cover the original goal, $G \setminus G' = \emptyset$. Note that our formulation of explanation generation problem captures the problem of generating explicable plans in [6]. At the same time, we also allow the goal of the problem to be changed. Compared to [1], we do not require the robot plan to be optimal in the updated model.

C. Explanation Generation

Our explanation generation problem can be considered as the problem of ***explicable goal augmentation***. Given a set of goal augmentation candidates or short as goal candidates, denoted as $\mathcal{G}$, which is a set of goal augmentations that the robot may choose to perform. In order to achieve the original goal, we consider the new goal as composed of the original goal and an additional goal that is in the goal candidates. That is, $G' \in \{G \cup \triangle G \mid \triangle G \in \mathcal{G}\}$. Then our explanation generation problem can be formulated as follows:
$G' = \text{argmax}_{G'}[\xi_{[\pi_{R'}:M^{H'}]}]$
where $G' \in \{G \cup \triangle G \mid \triangle G \in \mathcal{G}\}$, $M^{H'} = \langle D^H, I, G'\rangle$, and $\pi_{R'} \in \Pi_{R'}$ is a plan in $M^{R'} = \langle D^R, I, G'\rangle$. Then the

new goal G' can be used as the explanation–with the robot explaining to the human about this new goal.

Algorithm 1 Explanation Generation as Invese Planning

1: $\mathcal{E} = dict()$
2: **for all** $\triangle G \in \mathcal{G}$ **do**
3: $\quad G' = G \wedge \triangle G$ ▷ goal augmentation
4: $\quad M^{H'} = \langle D^H, I, G' \rangle$ ▷ update $M^{H'}$
5: $\quad M^{R'} = \langle D^R, I, G' \rangle$ ▷ update $M^{R'}$
6: $\quad$ **for all** $\pi_{R'} \in \Pi_{R'}$ **do**
7: $\quad\quad \epsilon = \xi_{[\pi_{R'}:M^{H'}]}$
8: $\quad\quad$ **if** $\epsilon == \xi_{max}$ **then return** G'
9: $\quad\quad$ **else**
10: $\quad\quad\quad \mathcal{E}[G'] = \epsilon$
return $\text{argmax}(\mathcal{E})$

The pseudocode of a native approach is presented in Algorithm 1. In line 6, instead of finding all possible plans, we can constrain to computing plans that are within a certain threshold of the optimal plans using the Fast-Downward planner [7]. Heuristics for this search will be discussed in future work.

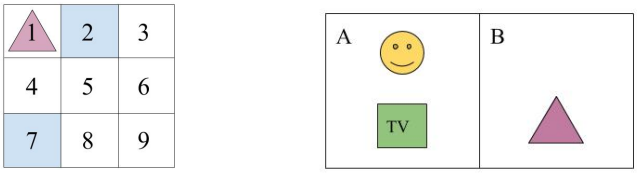
IV. Examples

A. Blocks World Exploration

A robot is exploring a block world shown in Figure 1(a). The original goal G is to explore $Cell_3$, $G = Cell_3$. The robot knows that there is a block in $Cell_2$, while the human does not. In this case, $\pi^*_R = (GO_4, GO_5, GO_6, GO_3)$, $\pi^*_H = (GO_2, GO_3)$, and $\xi_{[\pi_R:M_H]} = -3$. The candidate goals may include $T = \{Cell_1, Cell_5, Cell_7, Cell_9\}$. When the world is not observable to the human, generating explanations such as $Cell_2 = Blocked$ may not be always desirable as discussed. Instead, the robot searches for a new goal such that the plan becomes explicable to the human. In this example, the new goal that satisfies this requirement may be $G' = Cell_3 \wedge Cell_5$. Then, $\pi^*_{R'} = (GO_4, GO_5, GO_6, GO_3)$, $\pi^*_{H'} = (GO_4, GO_5, GO_6, GO_3)$, and $\xi_{[\pi_{R'}:M_{H'}]} = 0$. In this case, the robot would explain to the human that its goal is to explore $Cell_3$ and $Cell_5$ instead of telling the human about its original goal $Cell_3$. Note that the information about the blockage of $Cell_2$ is hidden.

B. House Chores

Some kids are watching TV programs in the living room as shown in Figure 1(b). The robot has a goal to clean the sofa on which the kids sit, $G = Clean\_Sofa$, and generates an optimal plan, $\pi^*_R = \{CLEAN\_SOFA\}$. However, the kids want to stay on the sofa for the TV programs as much as the robot wants to clean. So from their perspectives, they want the robot to wait and then clean the sofa, $\Pi^*_H = \{WAIT, CLEAN\_SOFA\}$. Even though the robot can generate the same plan as the kids wish, the waiting time is uncertain so that it may cause a significant cost (due to other tasks assigned). In this case, instead of providing explanations to persuade the kids to leave the sofa, the robot finds a new goal, $G' = Make\_Cake \wedge Clean\_Sofa$ and a corresponding optimal plan $\pi^*_{R'} = \{MAKE\_CAKE, CLEAN\_SOFA\}$, since making a cake will lure the kids to stay in the kitchen for the cake, while the robot can clean the sofa. Also, this plan is explicable to the kids, $\xi_{[\pi^*_{R'}:M_{H'}]} = 0$. Note that this scenario requires the robot to maintain a model of the kids.



(a) Block World (b) House Chores

Fig. 1: (a) A robot represented by the purple triangle is exploring the block world, where blocks are marked as blue. (b) There are kids watching TV programs in the living room, $ROOM_A$. A robot represented by the purple triangle is in the kitchen, $ROOM_B$.

V. Discussion

Based on the principle of minimally changing the human's knowledge to reduce the cognitive load, we proposed to generate explanations as explicable goal augmentations that result in explicable plans to a human while achieving the original goal at the same time. Although as an approach to explanation generation, it may also be used as a method to "hide" information from certain groups of people. This can be useful for non-collaborative or adversarial situations where information is considered private. Future work would also include providing heuristics to search for the new goal faster.

References

[1] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI*, 2017.

[2] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. Hierarchical expertise-level modeling for user specific robot-behavior explanations. *arXiv preprint arXiv:1802.06895*, 2018.

[3] Michelene TH Chi. Short-term memory limitations in children: Capacity or processing deficits? *Memory & Cognition*, 4(5):559–572, 1976.

[4] Anthony G Greenwald. Cognitive learning, cognitive response to persuasion, and attitude change. *Psychological foundations of attitudes*, pages 147–170, 1968.

[5] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang. AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775*, 2017.

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

[7] Malte Helmert. The fast downward planning system. *JAIR*, 26:191–246, 2006.