Human Centered Explanation for Goal Recognition

Abeer Alshehri[1,3] , **Tim Miller**[1] , **Mor Vered**[2] and **Hajar Alamri**[1,3]
[1]University of Melbourne, Melbourne, Australia
[2]Monash University, Melbourne, Australia
[3]King Khalid University, Abha, Saudi Arabia
{aalshehri,halamri}@student.unimelb.edu.au, tmiller@unimelb.edu.au, mor.vered@monash.edu

Abstract

This paper contributes to the ongoing work in XAI by exploring explainable goal recognition (GR) from a human-centered perspective. We described a human-centred study that informs a data-driven approach to understanding how people explain goal recognition tasks, with the future aim to build an explanatory model for GR. In our study, the participants attempt to infer an agent's goal given some observed behavior, and then provide explanations (why, why not, or both) for those inferred goals. Using thematic analysis process, we identified 11 codes from within 864 explanations of agents performing optimally, suboptimally, or irrationally in a Sokoban game variant. Interpreted with the existing theory of behavior explanation, we built a preliminary model for goal recognition explanations.

1 Introduction

Recently, there has been an increasing interest in constructing Explainable Artificial Intelligence (XAI) systems. Since humans are targeted by these systems, e.g., in medical diagnosis and autonomous driving [Szolovits *et al.*, 1988; Yaqoob *et al.*, 2019], researchers argue that building such systems should be based on social science–how humans explain behavior [De Graaf and Malle, 2017; Miller, 2018].

Given an agent's initial state, a set of candidate goals, and an incomplete sequence of observed actions, a goal recognition (GR) is the process of determining the most likely goal for that agent [Vered *et al.*, 2018]. Communicating goals of autonomous vehicles to end-users, for instance, help them to calibrate their trust to such a system [Shahrdar *et al.*, 2018]. We are motivated by the need to consider how humans explain others' goals, given their observed behavior. We believe that building 'artificial' explanations to mimic human explanations will lead to more natural and trusted explanations.

Consider the following instance of daily goal recognition: when a person observes another person walking towards a parked car, they assume that the first person will get into the car and drive away, hence freeing up a potentially coveted parking spot. Therefore the logical action to take would be to wait for that space to be freed. In such a scenario, the question arising is, "what made you wait?" One explanation may be that they saw you heading to the car carrying bags and predicted your goal based on that. From this explanation, we can learn something about how people may explain other people's goals. Investigating such explanations may provide further insight into the concepts humans use to reason about recognized goals. Namely, it reveals a noteworthy aspect of human cognition: How we infer others' mental states underlying their sparsely observed behavior.

In this paper, we contribute to the ongoing work in XAI by exploring explainable GR from a human-centered perspective. We aim to learn how people explain agents' goals and model the related concepts for such explanations. To obtain a formative understanding that can be used to design an explainable GR system, we conducted a case study in a controlled context as a first step. Our research question is: "what forms of explanations are given by humans for predicted goals in a goal recognition task?"

To answer that, we followed a data-driven approach to provide a model of GR explanations. We recruited 36 participants and asked them to infer an agent's goal from its observed actions in a game. We then asked them to explain the reasons behind their predictions by answering a follow-up question/s based on the condition they were in answering *why* for Condition 1, *why not* for Condition 2, or both *dual* for Condition 3. Using thematic analysis process, we identified 11 codes from within 864 explanations of agents performing optimally, suboptimally, or irrationally. Interpreted with the existing theory of behavior explanation, we present a human-centred model for GR explanations.

2 Background

This section highlights recent work that relates to explainable systems, goal recognition, and cognitive science.

2.1 Explainable AI

As AI systems are becoming increasingly complicated, the need for XAI arises to interpret "black-box" machine learning mechanisms and provide further understanding into autonomous agents' behavior. Consequently, XAI research can be categorized into two major branches: Data-driven and goal-driven XAI [Anjomshoae *et al.*, 2019]. In data-driven XAI, researchers aim to interpret black-box machine learning algorithms, such as how the given data led to a decision; A single-shot problem in which the model's output depends

on its input. Goal-driven XAI refers to autonomous agents (i.e., explainable agency). That is, an agent must be able to explain and reason about actions *sequence* that led to its decision and choice. As people tend to attribute mental components such as desires and beliefs to autonomous agents, building such systems based on social psychology and cognitive science helps people raise their trust [Miller, 2018].

Studies on XAI using social science mostly focus on single-shot machine learning mechanisms [Ren *et al.*, 2017; Adadi and Berrada, 2018; Ghosal *et al.*, 2018]. In this paper, we focus on goal-driven autonomous agents engaged in sequential tasks, explaining a decision as part of a sequential decision-making process made by an agent. A more typical scenario is that the agent performs an action based on the current observation, makes a new observation in which a newly performed action is conditioned on, and so on.

Several explanation frameworks have been proposed for Belief-Desire-Intention BDI agents [Cranefield *et al.*, 2017], reinforcement learning RL agents [Madumal *et al.*, 2019], and planning agents [Cashmore *et al.*, 2019]. These frameworks are mostly driven by goal-directed tasks over understanding the decisions of autonomous agents. Such approaches, however, have dealt with explaining certain actions based on a known goal. In GR systems, the explanation would be for the inverse problem: explaining an anticipated goal based on observed actions.

2.2 Goal Recognition Systems

The need to explain GR systems forces them to be made in a way that can be understood by people. Well-established algorithms can effectively recognize goals but hardly explain it. For instance, there are methods that have shown a great performance of labeling a sequence of actions with which goal they relate to [Ramírez and Geffner, 2010; Vered *et al.*, 2018]. However, explaining why the algorithm has derived that answer has not been investigated. Part of the problem lies in first understanding what such an explanation should contain.

Baker *et al.* [2009] proposed a computational model of how a human observer makes inferences of other's goals from their observed behaviour. The observer can infer goals by making a causal relationship between other's beliefs, goals and actions as rational planners solving a Markov decision process (MDP), a normative framework to model sequential tasks under uncertainty. Thus, this causal model of a sequential task can be used in an inverse direction to allow rational inferences of goals from observed behaviour. The agent preference is considered to be embedded within the agent's desired state (goal) as a value function that guides its decision under rationality principle (a goal-dependent cost function). That model is considered as a reasonable first approximation for a human mind to infer others' goal. However, this predictive model does not present the concepts that people use for the purpose of explaining predicted goals. Our proposed model deals with the challenge of explaining the goal of an observed sequence of actions performed by an acting agent.

2.3 Cognitive Science

Explaining behavior within an environment is a result of the cognitive process [Miller, 2018]. There are several proposed frameworks in social science on how people explain behavior [Heider, 1982; Kashima *et al.*, 1998; Malle and Tate, 2006]. Malle [2006] defines a comprehensive model of explaining behavior called the *folk theory of mind and behaviour*. It is an attribution of human behavior applying everyday terms such as desire, belief, valuing, and intention. Humans tend to use these concepts to understand and explain their own and others' behaviors. By forming an explanation in this conceptual framework, it can identify different modes of behavior explanations and their cognitive processes. Malle represents the assumptions and distinctions that humans typically make between intentional and unintentional behavior. Thus his model shows two contrasting modes of explanations. The primary mode of explaining intentional behavior is reasoning over key mental components—the reasons behind this deliberate act. It is based on the rationality principle: intentional agents are expected to act efficiently towards achieving their desires, given their beliefs and values. The second mode is explaining unintentional behavior by referring to causes such as habitual or physical phenomena, for example, the causes behind an accidental or unintentional act.

For this framework, the goal/desire for intending certain actions is known beforehand. Citing the goal is considered as a way to explain behavior [Malle and Tate, 2006]. However, the inverse problem of explaining goal recognition has not been studied as far as we know. We aim to take insights from Malle's framework [Malle and Tate, 2006] to solve the inverse problem of explaining behavior in which the goal/desire is unknown.

3 Methodology

Next we describe our research methodology. Our aim is to identify: 1) the concepts people use to explain goal recognition; 2) any patterned relationships within these concepts; and 3) concept frequency across different rationality levels.

3.1 Research Design

The research design is based on a data-driven approach using thematic analysis [Braun and Clarke, 2006] as a method of analyzing data.

The data analysis was divided into six phases: familiarizing with the collected data, developing codes, sorting the different codes into potential themes, reviewing themes, defining and naming themes, and writing up the report [Braun and Clarke, 2006]. The collected data was re-read several times, resulting in data immersion before proceeding to the coding phase. The coding involved an inductive process identify the basic concepts, such as facts and preferences that are relevant to the research question of explaining human behavior. The codes were then sorted into defined themes based on similarity. In the context of this study, proposed themes are associated with our research topic of explaining human behavior in goal recognition scenarios. After establishing a set of candidate themes, the refinement process looked for internal homogeneity—a cohesive pattern of the candidate themes to reflect what was observable in the overall data set. Moreover, relationships, links, and distinctions between themes were identified. The next step involved naming and describing the

themes and finally demonstrating the thematic elements with examples.

3.2 Modified Sokoban Game

We conducted an online Wizard of Oz experiment and built a modified version of the game *Sokoban* (Figure 1). Sokoban is a puzzle game where a player/agent moves boxes around in a warehouse to deliver them to storage locations. We modified the game rules to allow the player to push more than one box simultaneously. Thus, it is not just a navigational task, but a strategic one with multiple goals in which the player aims to minimize the number of steps.

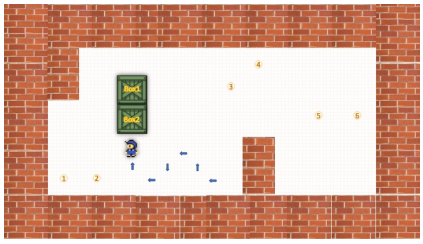


Figure 1: Example screenshot of the game

We described a participant/observer's world view of the game scenarios as a Markov decision process (MDP). Our description was based on three assumptions: (1) the observation sequence is complete and fully observable; (2) the observer reasoning is based on their internal world state representation, that does not necessarily match the observed player/agent; and (3) the player's preferences are not observable.

Formally, we define $\acute{S}_t$ as the world state at time t that consists of all objects on the map (walls, boxes), constraints, goal states, and agent state. $G_t \subseteq \acute{S}_t$ is a finite set of possible goals. A_t is a finite set of possible actions at time t. The player can take one of two actions: (1) a *move* action (up, down, right, left); and (2) a *push* action (up, down, right, left). All action effects are deterministic.

There were different goal representations based on the game version (Table 1). In versions 1 and 2, the player can push one box at a time, whereas in version 3, it can push two boxes at a time to a goal location/s. The goal recognition problems used a number of competing goal hypotheses: different possible plans to achieve a goal, two sequential goals with interleaved plans to achieve them, or failed/unsolvable plans.

Version	Player Task
Game 1	Deliver one box to one of three possible goal locations; push one box at a time.
Game 2	Deliver two boxes to two of four possible goal locations; push one box at a time.
Game 3	Deliver two boxes to two of six possible goal locations; push multiple boxes at a time.

Table 1: Player game versions.

We varied scenarios to present different *rationality levels*, either rational (optimal or suboptimal) behaviors, or irrational. For optimal behaviour, we assumed that observers would form a simple notion of optimal behavior, that is, the player will take the shortest plan towards a goal. In the suboptimal behavior, the player either chooses a longer path towards a goal (suboptimal plan) or deviates from a rational action in an observed sequence of a particular goal plan (e.g. the agent's goal may have changed). We also presented some irrational behaviors in which the player failed to achieve the task (i.e., getting stuck in a dead-end state) to find out how those would be explained. Having different levels of rationality helps to see whether observers explain suboptimal behavior differently from optimal behavior.

3.3 Data Collection

We recruited 36 participants (22M/14F) – 12 in each condition – who resided in the USA and were aged between 20 and 65. Participants were recruited via Amazon Mechanical Turk. Participants were paid $6.50 for completing the task and a bonus of $3.50 for giving more 'thoughtful' answers. We gathered and coded a total of 864 explanation responses. All collected data is text-based.

Task details. The participants' task is divided into the following phases:

1. Watch an instructional video to introduce the participant to the task and game rules.
2. For 18 different scenarios of three games (six scenarios per game):
   (a) Watch a scenario (video clip) in which a player tries to achieve the task (Table 1).
   (b) After watching the observed actions sequence (plan's completion percentage *min*=0.25%, *median*=0.53%, and *max*=0.83%), participants were asked to predict which goal location the player was trying to get to, and, accordingly, assign a likelihood (with one as the least likely and five as the most likely) of each goal.
   (c) Give reasons for their prediction, answering specific questions based on the condition they are in.

Each participant was assigned into one of three conditions: 'why' condition, 'why-not' condition, and 'dual' condition

- 'Why' condition: participants were asked: "Explain *why* have you rated that/those goal(s) as the most likely?"
- 'Why-not' condition: participants were asked: "Explain *why you have not* rated that/those goal(s) as the most likely?"
- 'Dual' condition: participants were asked *both* questions of *why* and *why you have not* in that order.

	#Questions per game				
Conditions	G1	G2	G3	Participants	Explanations
Why	6	6	6	12	216
Why-not	6	6	6	12	216
Dual	12	12	12	12	432

Table 2: Data sources

We collected data for the first and second conditions to analyze the differences between *why* and *why not*, and the third condition to analyze how people answer *why not* if they have already answered *why*, and how the answer of *why* differs if they know there is a *why not*. We also used participants' open-ended explanations to identify better the concepts they use to explain the player's predicted goal. The word count of given answers within the data set is between 1 and 98 words (M=22.96, SD=15.52) for the first condition, 3 and 81 words (M=26.57, SD=15.63) for the second condition, and 1 and 64 words (M=20.38, SD=11.65) for the third condition.

With three different game versions, six scenarios by game, and 12 participants per condition, the result is 864 different explanations (see Table 2). The map configuration and level of rationality vary between each scenario to get different possible explanations.

4 Results

In this section, we present our results: a model of explainable goal recognition. We highlight and analyze the key concepts of a goal recognition explanation, relationships between these concepts, and the frequency of concepts occurring across different rationality levels. We also evaluate the mode of explanations given for each condition.

4.1 Explanation Goal Recognition Model

Figure 2 presents our model of explaining goal recognition derived from the study. The proposed explanation model is based on Malle and Tate [2006]'s framework for explaining behaviour. First, as the problem of recognition activates counterfactual thinking [Epstude and Roese, 2008], counterfactual cases (facts and preferences) are presented as a part of mental states. The reasoning process among possible goals to answer why/why not questions is inherently a contrastive process. Namely, causal inferences indicate an agent's goal, given its current beliefs and observed behavior. By integrating that with contrastive reasoning among G_t, it highlights salient differences between goal plans Π_t and uses that to explain G_t (intended-goal hypotheses and counterfactual-goal hypotheses). Secondly, we break down the fact concept to include an *observational marker*, assuming that the agent can calculate them as a part of its belief state. Besides, preference is implemented separately as a mental state rather than a part from the goal state following [Malle and Tate, 2006], since it could be possible for a system to satisfy the goal without satisfying the preference. [Visser *et al.*, 2016]. The observers reason about the closest/effective path, assuming that the agent's preference is the nearest goal location. As people prefer reasoning over preferences in the first place to explain possible goals [Winikoff *et al.*, 2018], we implemented different levels of rationality besides an unobservable preference state to ensure having rich explanations that are unbounded by a preference state.

4.2 Codes

Table 3 shows the codes, categories, and descriptions. Looking at the observer view of the world, we can see the key components of the explanation goal recognition model (figure 2). We coded all information across three different conditions for three-game versions. Explanations were formed as combinations of different codes based on the context.

As explanations come in terms of the mental states and intentions, we propose a new concept that is part of a belief mental state.

Observational marker. An *observational Marker* is an action or sequence of actions executed at some point along some valid plan to achieve a particular goal set. Given the goal recognition problem (refer to Section 3.2), the observer infers goals by finding evidence from achieved actions in O_t of some *rational* plan π_t. These action/s are considered an observational marker that is true along any π_t towards achieving the intended-goal hypotheses.

It is a different concept from proposed landmarks in planning [Hoffmann *et al.*, 2004]. The difference is that a landmark must be executed at some point along *all* valid plans to achieve a particular goal, while the observational marker is along *any* valid plan toward a particular goal.

Observers contrast among Π_t and explain G_t in terms of observational markers that are presented in π_t of the intended-goal hypotheses, but not of the counterfactual-goal hypotheses. The following are special cases of observational markers:

- If there is only one valid plan π_t to intended-goal hypotheses, then observers mostly suffice to explain using only observational markers.
- If there is a deviation at the last moment from a rational action–a deviation from a rational sequence of O_t toward a particular goal set, then that last deviated action/s is considered as an observational marker to explain new intended-goal hypotheses.

4.3 Relationships

The model captures the causal relationships of how different mental states cause goals using the principle of rational planning/rational action. People tend to consider future plans/actions that are easily inferred from mental states [Nakahashi and Yamada, 2018]. Their explanations are guided either by simulating or inferring the causal link between the world's current state and their hypothesized goal [Malle and Tate, 2006].

In a sequential task, people tend to infer the causal effect of interest (salience conditions). Causal relations assist conditional reasoning based on the link between mental states (factual and counterfactual conditions) and a goal set. People look to contrary conditions that contrast with current factual conditions when they form plans to reach the anticipated goal [Lombrozo, 2012]. Thus, contrastive explanations are the method of inferring the constraints from the mental states towards achieving certain goals. This notion of constraint is responsive to salience conditions from current states towards each possible goal. Thus, if C and D were to cause $g_x \in G$ from the following sequence of actions $\{C, D, ..., B, g_x\}$, then C and D are salience sequential conditions that are satisfied in π_{g_x}, not other's goal plan. For example, consider an explanation from the data corpus, "because the box was pushed down towards goals two and three". Here, the salience condition is the action "box pushed down" that is satisfied

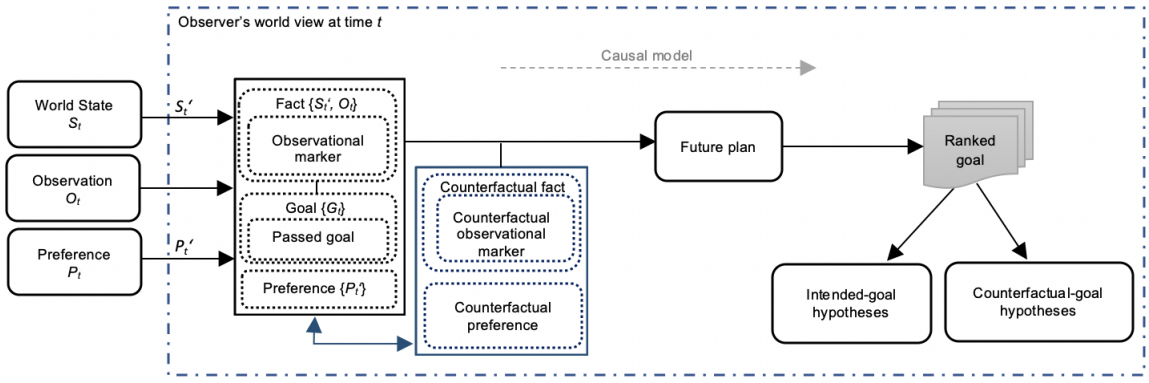


Figure 2: Explanation Goal Recognition Model

Code	Category	Description
Fact	Belief	A world state, observations and possible actions and its effects
Observational marker	Belief	An action or sequence of actions that is true at some point along any valid plan to the intended goal hypotheses
Preference	Valuing	Positive value
Goal	Desire	A set of possible goals
Passed goal	Desire	Goal state that has been passed
Counterfactual fact	Belief	Fact that is contrary to the factual world
Counterfactual obs. marker	Belief	An action or sequence of actions that is true at some point along a valid plan to the counterfactual goal hypotheses
Counterfactual preference	Valuing	Negative value
Future plan	Intention	Formed plan from current state towards goals
Intended goal hypotheses	Desire	Predicted goal of higher likelihood
Counterfactual goal hypotheses	Desire	Predicted goal of lower likelihood

Table 3: Code description

in the plan of goal locations two and three. By contrasting different goal plans, pushing the box down would be a constraint from reaching goal location one, which has been used to explain the predicted goals. Clearly, the given explanations only state the most significant changes required to adjust the model's prediction.

As we varied goal representations to have two goals (refer to Table 1), we found that observers' goal traces were either conditionally dependent or independent of each other. For the former, they were giving an explanation for achieving a goal sequence dependently on one another; That implies a temporal ordering—conditioning one's plan over another. In the latter, they were giving an explanation for a goal sequence independently of one another, that is by conditioning each goal of the sequence only on the current mental states. Table 3 shows the codes, categories, and descriptions. Looking at the observer view of the world, we can see the key components of the explanation goal recognition model (figure 2). We coded all information across three different conditions for three-game versions. Explanations were formed as combinations of different codes based on the context.

4.4 Code Occurrence across Rationality Levels

We analyzed the given explanations across all game versions, considering the different levels of rationality implemented. Figure 3 shows the frequencies of 10 identified codes. Goals are always mentioned, so have thus, it has been omitted. Overall, observers have shown to be strongly reliant on both observational markers and counterfactual observational markers when explaining their predictions. They were introduced as a part of their explanations in most scenarios. We believe that these two concepts can even help to predict and explain cases where there were few observations. Also, future plans were used as a part of explanations frequently. In fact, explaining using future plans (intentions) inherently involves a reference to a personal goal.

Optimal Behaviour Explanations We believe that people explain mostly with regard to optimality. The evidence comes from the theory of mind [Csibra, 2017], where people show a tendency to use their internal reasoning, based on their understanding, for the purpose of optimization. Based on decision theory, a rational agent models uncertainty to make a decision via a clear preference—a choice with maximum utility [Weirich, 2008]. Observers captured the optimality principle in that sense; they chose the shortest path from the current state to the goal states. They assumed that the agent reasons and contrasts over goal plans to choose the one with the lowest cost for the best explanation. Miller [2018] states that it is the prerequisite of optimality that not only guarantees the validity of future expected plans, but it also shows a preference, compared to other alternatives.

Observers first placed reliance on the observational marker concept and combined it when possible to the preference/counterfactual preference state; that is, for the 'best' explanation when they have more than one predicted goal and

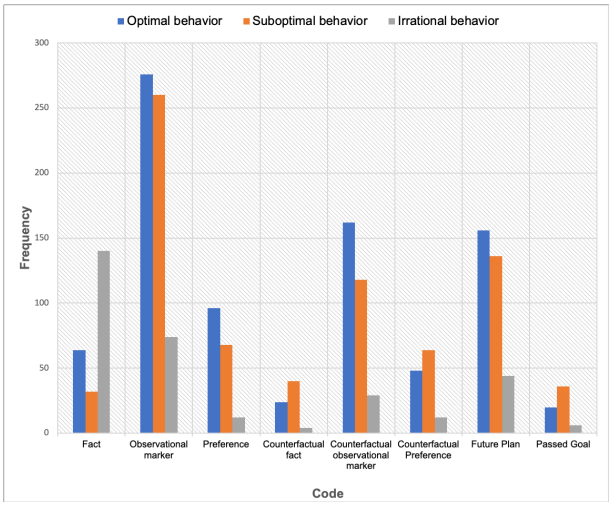


Figure 3: Code occurrence across different rationality levels

contrast between them.

Additionally, from a rationality perspective, observers assumed that the player would not go back to any goal location once it is passed, and here is where they used *passed goal* concept as a part of explanations.

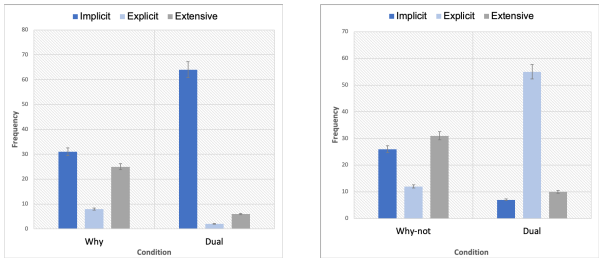
Suboptimal Behaviour Explanations Because we introduce suboptimal behavior in the instructional video stage, observers seem to assume that sub-optimal behaviour was acceptable; e.g. a player taking a longer route or changing its goal. In such cases, participants tend to explain facts and counterfactual states, that is, to reason about the player's abilities from its current state or give information about the game rules. In some cases, it comes in the form of incorrect statements from a misunderstanding of the given scenario. When players revised their beliefs, observers explained only in terms of the deviated action/s (revised beliefs), and that action was considered an observational marker of the new goal. In addition, as the purpose of preference is to guide the agent to make a rational decision, observers still reason about preference and assume to be the location's cost rather than the path cost. For example, from the data corpus: "although it became so close to goal 3, he changed his path to goal 1, which makes it more preferable than 3".

Irrational Behaviour Explanations Looking into irrational scenario explanations helps to see how invested people are in the rationality principle to explain irrational behavior. Agents could sometimes look irrational to observers, and still, they need an explanation. From the rationality principle, people explain other's desires even if there is no intention to it [Malle and Tate, 2006]. From that, we found observers still predict the player's intended goal for the case of being stuck in a dead-end state (unable to reach that goal). For instance, from data corpus: "pushed themselves into a dead end, but most likely he meant to go for position 1 because they overshot position 2 and 3". Malle and Tate [2006] state that explaining beliefs comes in terms of other beliefs that support them. Observers explain the failure in terms of current facts that cause failure (i.e., game rules of being stuck in a dead-end state). We can see the use of other concepts in the explanations, and they are presented in the game versions two and three, where only one box is stuck. Thus, observers still explain the predicted goal location of the non-stuck box.

4.5 Modes of Explanation

Lipton [1990] argues that why-questions ask for a contrastive explanation; that is, answering why A is in the form of "why A instead of B?", where B is some foil case that did not happen. A good answer to such a question will be in terms of a causal difference between A and B: what needs to be changed for the foil to happen.

In goal recognition, event A represents intended-goal hypotheses, whereas event B represents counterfactual-goal hypotheses. From the collected data, we classified the given contrastive explanations into three modes: *implicit* in which observers explain only the facts (intended-goal hypotheses); *explicit* in which observers explain only the foils (counterfactual-goal hypotheses); and *extensive* in which observers explain both facts and foils.



(a) Why question (b) Why-not question

Figure 4: Frequency of explanation modes for different questions

Figure 4 shows the difference between conditions. Observers tended to give more implicit explanations in the 'dual' condition — answering why-questions when they already answered why-not—compared to 'why' condition (Figure 4a). Observers gave more explicit explanations—answering why-not when they already answered why—compared to 'why-not' condition (Figure 4b). The contrastive nature of explanations is clear in the 'dual' condition where observers can distinguish between two types of questions. Clearly, a natural explanation from a GR system will explain things differently, depending on which questions have already been answered.

5 Conclusion

This paper has introduced a preliminary model for a GR system. We have identified the key concepts for goal recognition explanations using thematic analysis process. GR systems that need to be explainable will benefit from such a model in providing more natural and intuitive explanations. In future work, we aim to evaluate the model in both collaborative and non-collaborative settings. Further evaluation can be done by having a partial observation sequence—a subsequence of observed actions—which may introduce different concepts.

References

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[Anjomshoae *et al.*, 2019] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[Baker *et al.*, 2009] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

[Braun and Clarke, 2006] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[Cashmore *et al.*, 2019] Michael Cashmore, Anna Collins, Benjamin Krarup, Senka Krivic, Daniele Magazzeni, and David Smith. Towards explainable ai planning as a service. *arXiv preprint arXiv:1908.05059*, 2019.

[Cranefield *et al.*, 2017] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. In *IJCAI*, pages 178–184, 2017.

[Csibra, 2017] Gergely Csibra. Cognitive science: modelling theory of mind. *Nature Human Behaviour*, 1(4):1–1, 2017.

[De Graaf and Malle, 2017] Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

[Epstude and Roese, 2008] Kai Epstude and Neal J Roese. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192, 2008.

[Ghosal *et al.*, 2018] Sambuddha Ghosal, David Blystone, Asheesh K Singh, Baskar Ganapathysubramanian, Arti Singh, and Soumik Sarkar. An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, 115(18):4613–4618, 2018.

[Heider, 1982] Fritz Heider. *The psychology of interpersonal relations*. Psychology Press, 1982.

[Hoffmann *et al.*, 2004] Jörg Hoffmann, Julie Porteous, and Laura Sebastia. Ordered landmarks in planning. *Journal of Artificial Intelligence Research*, 22:215–278, 2004.

[Kashima *et al.*, 1998] Yoshihisa Kashima, Allison McKintyre, and Paul Clifford. The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1(3):289–313, 1998.

[Lipton, 1990] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[Lombrozo, 2012] Tania Lombrozo. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276, 2012.

[Madumal *et al.*, 2019] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. *arXiv preprint arXiv:1905.10958*, 2019.

[Malle and Tate, 2006] Bertram F Malle and Chuck Tate. Explaining the past, predicting the future. *Judgments over time: The interplay of thoughts, feelings, and behaviors*, pages 182–209, 2006.

[Miller, 2018] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

[Nakahashi and Yamada, 2018] Ryo Nakahashi and Seiji Yamada. Modeling human inference of others' intentions in complex situations with plan predictability bias. *arXiv preprint arXiv:1805.06248*, 2018.

[Ramírez and Geffner, 2010] Miguel Ramírez and Hector Geffner. Probabilistic plan recognition using off-the-shelf classical planners. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[Ren *et al.*, 2017] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. Social collaborative viewpoint regression with explainable recommendations. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 485–494. ACM, 2017.

[Shahrdar *et al.*, 2018] Shervin Shahrdar, Luiza Menezes, and Mehrdad Nojoumian. A survey on trust in autonomous systems. In *Science and Information Conference*, pages 368–386. Springer, 2018.

[Szolovits *et al.*, 1988] Peter Szolovits, Ramesh S Patil, and William B Schwartz. Artificial intelligence in medical diagnosis. *Annals of internal medicine*, 108(1):80–87, 1988.

[Vered *et al.*, 2018] Mor Vered, Ramon Fraga Pereira, Maurício Cecílio Magnaguagno, Gal A Kaminka, and Felipe Meneguzzi. Towards online goal recognition combining goal mirroring and landmarks. In *AAMAS*, pages 2112–2114, 2018.

[Visser *et al.*, 2016] Simeon Visser, John Thangarajah, James Harland, and Frank Dignum. Preference-based reasoning in bdi agent systems. *Autonomous agents and multi-agent systems*, 30(2):291–330, 2016.

[Weirich, 2008] Paul Weirich. Causal decision theory. 2008.

[Winikoff *et al.*, 2018] Michael Winikoff, Virginia Dignum, and Frank Dignum. Why bad coffee? explaining agent plans with valuings. In *International Conference on Computer Safety, Reliability, and Security*, pages 521–534. Springer, 2018.

[Yaqoob *et al.*, 2019] Ibrar Yaqoob, Latif U Khan, SM Ahsan Kazmi, Muhammad Imran, Nadra Guizani, and Choong Seon Hong. Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. *IEEE Network*, 2019.