Journal of Artificial Intelligence Research 82 (2025) 2535-2594 Submitted 09/2024; published 04/2025

Towards Explainable Goal Recognition Using Weight of Evidence (WoE): A Human-Centered Approach

Abeer Alshehri AALSHEHRI@STUDENT.UNIMELB.EDU.AU
School of Computing and Information Systems
The University of Melbourne, Melbourne, Australia
Department of Computer Science and Information Systems
King Khalid University, Abha, Saudi Arabia

Amal Abdulrahman AMAL.ABDULRAHMAN@MQ.EDU.AU
School of Computing
Macquarie University, Sydney, Australia

Hajar Alamri HAJMOHEMMAD@KKU.EDU.SA
Department of Computer Science and Information Systems
King Khalid University, Abha, Saudi Arabia

Tim Miller TIMOTHY.MILLER@UQ.EDU.AU
School of Electrical Engineering and Computer Science
The University of Queensland, Brisbane, Australia

Mor Vered MOR.VERED@MONASH.EDU
School of Computing and Information Systems
Monash University, Melbourne, Australia

Abstract

Goal Recognition (GR) involves inferring an agent's unobserved goal from a sequence of observations. This is a critical problem in AI with diverse applications. Traditionally, GR has been addressed using 'inference to the best explanation' or abduction, where hypotheses about the agent's goals are generated as the most plausible explanations for observed behavior. Alternatively, some approaches enhance interpretability by ensuring that an agent's behavior aligns with an observer's expectations or by making the reasoning behind decisions more transparent. In this work, we tackle a different challenge: explaining the GR process in a way that is comprehensible to humans. We introduce and evaluate an explainable model for goal recognition (GR) agents, grounded in the theoretical framework and the cognitive processes underlying the explanation of human behavior. Drawing on insights from two human-agent studies, we propose a conceptual framework for human-centered explanations of GR. Using this framework, we develop the *eXplainable Goal Recognition* (XGR) model, which generates explanations for both *why* and *why not* questions. We evaluate the model computationally across eight GR benchmarks and through three user studies. The first study assesses the efficiency of generating human-like explanations within the Sokoban game domain, the second examines perceived explainability in the same domain, and the third evaluates the model's effectiveness in aiding decision-making in illegal fishing detection. Results demonstrate that the XGR model significantly enhances user understanding, trust, and decision-making compared to baseline models, underscoring its potential to improve human-agent collaboration.

1. Introduction

Goal Recognition (GR) is the problem of predicting an agent's intent by observing its behavior. The task of GR has numerous potential and practical applications, such as smart homes (Hegde & Kenchannavar, 2019) and workplace safety (Inam et al., 2018), among others (Singh et al., 2020; Wayllace et al., 2020). Research on GR uses different inference techniques to predict the ultimate goals of the agents being observed. It advances with increasingly complex domain models and better approaches. However, understanding and fostering human trust in these systems is challenging due to their lack of explainability. This becomes particularly crucial in safety-critical applications like social care, military planning, and medical support, where the system's decisions can have major consequences. Systems must be capable of explaining the decisions made and communicating their decisions in a way that is understandable to people (Masters & Vered, 2021; Meneguzzi & Pereira, 2021; Van-Horenbeke & Peer, 2021).

The vast majority of work has focused on improving the explicability of agent behavior (Hanna et al., 2021; Hu et al., 2021; Sohrabi et al., 2016; Vered et al., 2016; Yolanda et al., 2015). This typically involves making the behavior more understandable to observers, either by aligning it with their expectations or ensuring the interpretability of the inference process. In shifting the focus from interpretability to justification, the goal is to develop an explainable GR agent that provides context and rationale for each predicted goal, rather than merely making the inference process interpretable.

This paper contributes to ongoing work on eXplainable AI (XAI) by developing an explainable GR model from a human-centered perspective. Studies in social psychology indicate that people use the same conceptual framework that they apply to humans to explain artificial agents' behavior, and they also expect artificial agents to adopt this framework (De Graaf & Malle, 2017). Therefore, our approach is to provide comprehensible explanations that align with how people rationalize outcomes and interpret information. This effort aims to bridge the gap between mere prediction and understandable explanation.

There are two main contributions of this work. First, we propose a conceptual framework for the explanation process in GR tasks and identify key concepts people use when reasoning about goal prediction. Using a bottom-up approach, the framework is derived from an analysis of human explanations of recognition tasks. We study this in two different domains to increase the generalizability of our model: Sokoban and StarCraft games. We examine the frequency, sequence, and relationships between the basic components of these explanations. Using the thematic analysis process, we identified 11 key reasoning concepts that emerged from analyzing 864 explanations of agent behavior across various scenarios. These concepts encapsulate common themes and patterns of reasoning within the provided explanations. Incorporating insights from the folk theory of mind and behavior (Malle, 2006a), we propose a human-centered model for GR explanations.

Our second contribution is building an *eXplainable Goal Recognition* (XGR) model based on the proposed conceptual framework. The model generates explanations for GR agents using the information theory concept of Weight of Evidence (WoE) (Good, 1985; Melis et al., 2021). We define the problem of explanation selection using the main concept from our conceptual model, which we call an *observational marker*, i.e., the observation with the highest WoE. We computationally evaluate the XGR model in eight GR benchmark domains

(Vered et al., 2018). We conducted three user studies to evaluate our model's performance in different aspects: 1) Efficiency in Generating Human-Like Explanations: This study focused on assessing how well the model produces explanations that resemble those generated by humans, specifically within the Sokoban game domain; 2) Perceived Explainability: This study examined how users perceive the model's ability to explain its actions and decisions, also within the Sokoban game domain. 3) Effectiveness in Supporting Decision-Making: This study evaluated the model's effectiveness in supporting users' decision-making process in the domain of illegal fishing detection. In the first study, our model aligns with human explanations in more than 73% scenarios. In the second and third studies, our model outperforms the tested baselines.

Part of this paper was published at the International Conference on Automated Planning and Scheduling (ICAPS) (Alshehri et al., 2023), where we presented our XGR model and its evaluation through the first two user studies. In this work, we introduce the conceptual framework for GR explanations, provide further details on our XGR model, and extend its evaluation to the context of decision-making support by conducting a third user study.

The structure of the paper is as follows. Section 2 reviews the related work on explainability in GR and human behavior explanation; Section 3 presents the human-agent experiment to build the conceptual framework; Section 4 provides the necessary background required to follow up with the proposed XGR model; Section 5 presents the XGR model; Section 6 describes experiments that evaluate the model; We then conclude with a summary and opportunities for further research in Section 7.

2. Preliminaries

This section introduces relevant background and notation.

2.1 Planning

Planning aims to find a sequence of actions given an environment model, a current situation, and the goal to be achieved (Geffner & Bonet, 2022). The concept of planning is key to understanding GR algorithms that use planners in the recognition process. We build upon the following planning problem definition as defined in R. Pereira et al. (2017):

Definition 2.1. A planning task is represented by a triple $\langle \Xi, I, g \rangle$, in which $\Xi = \langle F, A \rangle$ is a planning domain definition that consists of a finite set F of facts that defines the state of the world, and a finite set A of actions; I is the initial state, and g is the goal state. A solution to a planning task is a plan π that reaches a goal state g from the initial state I by following transitions defined in Ξ. Since actions have an associated cost, we assume that this cost is 1 for all actions.

2.2 Goal Recognition (GR)

Goal recognition (GR) involves identifying an agent's goal by observing its interactions within an environment (Sukthankar et al., 2014). We consider the GR definition as defined by Shvo and McIlraith (2020).

Definition 2.2. A goal recognition problem is a tuple $\langle \Xi, I, G, O \rangle$, in which $\Xi = \langle F, A \rangle$ is a planning domain definition where F and A are sets of facts and actions, respectively; I is the initial state; $G = \{g_1, g_2, ..., g_m\}$ is the goals set, and $O = \langle o_1, o_2, ..., o_n \rangle$ is a sequence of observations such that each o_i is a pair (α_i, ϕ_i) composed of an observed action $\alpha_i \in A$ and a fact set that represent the state $\phi_i \subseteq F$. A solution to a GR problem is a probability distribution over G giving the corresponding likelihood of each goal, i.e. the posterior probability $P(g_j \mid O)$ for each $g_j \in G$. The most likely goal is the one whose generated plan 'best satisfies' the observations.

2.2.1 The Mirroring GR Algorithm

We focus on providing explanations for the output of the *Mirroring* GR algorithm (Kaminka et al., 2018; Vered & Kaminka, 2017). However, our approach is agnostic of the underlying GR algorithm and will work for any GR algorithm that fits Definition 2.2. The *Mirroring* algorithm is inspired by people's ability to perform online GR, originating from the brain's mirror neuron system, which is responsible for matching the observation and execution of actions (Rizzolatti, 2005). The approach falls under the *plan recognition as planning* GR approaches (Masters & Vered, 2021; Ramırez & Geffner, 2010) and uses a planner within the recognition process to compute alternative plans.

Specifically, the mirroring algorithm uses a planner to compute optimal plans from an initial state I to each goal $g_j \in G$ and to compute *suffix* plans from the last observation $o_i \in O$ to each goal $g_j \in G$. Observations are processed and evaluated incrementally. These *suffix* plans are then concatenated with a *prefix* plan (the observation sequence O at time step t) to generate new plan hypotheses. The algorithm subsequently provides a likelihood distribution, that is, posterior probabilities $P(g_j \mid O)$ for each $g_j \in G$ by evaluating which of the generated plans, incorporating observations O, best matches the optimal plan. The running example (see Section 5.1) illustrates The Mirroring approach to solving a GR task.

2.3 Weight of Evidence (WoE)

The principle of rational action (Hempel, 1961) states that people explain goal hypotheses by assessing the extent to which each observed action contributes to a specific goal hypothesis over others. Building on this idea, Bertossi (2020) defines a causal explanation as the set of features most responsible for an outcome. By incorporating this approach, we model our explanation framework using the Weight of Evidence (WoE) concept.

Weight of Evidence (WoE) is a statistical concept used to describe the effects of variables in prediction models (Good, 1985). It is defined in terms of log-odds (see supplementary material) to measure the strength of evidence e supporting a hypothesis h against an alternative hypothesis h', conditioned on additional information c. Assuming uniform prior probabilities[1], WoE is expressed as:

$$woe(h/h' : e \mid c) = \log \frac{P(h \mid e, c)}{P(h' \mid e, c)} \tag{1}$$

Melis et al. (2021) propose a framework based on WoE for explaining machine learning classification problems, arguing that it aligns with how people naturally explain phenomena to

1. The derivation of the formula for non-uniform priors is provided in Appendix B.

one another (Miller, 2019a). They found that WoE effectively captures contrastive statements, such as evidence for or against a particular outcome. This helps answer questions like why a goal g is predicted, why not goal g', and what should have happened instead if the goal is g'. We adopt this concept and apply it to GR problems.

3. Related Work

3.1 Goal Recognition and Explainability

Goal recognition (GR) involves identifying an agent's unobserved goal based on a sequence of observations. Various approaches exist to address the GR problem. Common approaches include library-based GR algorithms, which use specialized plan recognition libraries to represent all known approaches for achieving known goals (Sukthankar et al., 2014); model-based GR algorithms (Ramırez & Geffner, 2010; Santos et al., 2021; Sohrabi et al., 2016; Vered et al., 2016), where GR agents leverage domain knowledge through planners to generate the necessary plans for achieving a goal (Masters & Vered, 2021); and machine learning GR approaches that rely on large training datasets from which algorithms learn domain constraints (Amado et al., 2022; Chiari et al., 2023; Fitzpatrick et al., 2021; Meneguzzi & Pereira, 2021; Min et al., 2014; R. F. Pereira et al., 2019). Well-established algorithms have shown high performance in labeling action sequences with corresponding goals (R. F. Pereira et al., 2020; Ramırez & Geffner, 2010; Vered et al., 2018). However, these approaches focus primarily on identifying goals effectively rather than explicitly providing human-understandable justifications for their conclusions. As a result, the explainability of goal recognition decisions remains underexplored.

Previous work has focused on explaining GR in the form of answering the question: What goal is the agent trying to achieve? A long line of work has suggested explaining goal inference, which is a form of 'inference to the best explanation', also called abduction (Baker et al., 2008; Baker, 2012; Blokpoel et al., 2013; Van Rooij et al., 2008; Zhi-Xuan et al., 2020). This process formulates hypotheses about the agent's goals, which are identified as the most plausible explanations for the observed actions. That would assist in making sense of actions and attributing appropriate goals or intentions to them, enhancing our understanding of the agent's behavior. Another approach is improving the explicability of agent behavior (Cohen & Galescu, 2023; Farrell & Ware, 2020; Hanna et al., 2021; Hu et al., 2021; Sohrabi et al., 2016; Vered et al., 2016; Yolanda et al., 2015). This involves ensuring that the behavior of the agent is self-explanatory to an observer by either aligning its actions with the observer's expectations or making the reasoning behind its decisions interpretable. These approaches often assume optimal or simplified sub-optimal actions, neglecting the inherent challenges in agent planning. Keren et al. (2014) introduced the Goal Recognition Design (GRD) approach, which facilitates the process of inferring an agent's goals. This approach aims to analyze and redesign the underlying domain environment to ensure early and accurate detection of the agent's objectives.

However, previous approaches assume that the agent's behavior or domain is controlled to make its actions explicable. In this work, we address a different problem: explaining the GR process in a way that is understandable to humans. In complex and dynamic environments where agent behavior may not neatly align with human expectations or optimal action models, understanding the reasoning behind goal predictions is crucial for building

trust in GR systems. There is a need for an explanation model that justifies the GR output, ensuring that the reasons behind goal predictions are clear and understandable. Instead of merely making the inference process transparent by controlling the domain or agent's actions, a model should provide context and rationale for each predicted goal. This includes accounting for sub-optimal behavior that might arise due to planning difficulties or environmental constraints (Masters & Sardina, 2021), thereby offering a more nuanced and realistic understanding of agent actions.

Additionally, fostering a solid understanding of an agent's behavior presents a significant challenge for decision-makers. Human-AI team performance is influenced by scenarios where the AI system provides predictions while humans maintain the final decision-making authority. GR systems play an essential role in this context by accurately predicting and interpreting user intentions, which inform and guide subsequent actions (Brewitt et al., 2021; Jamakatel et al., 2023; Ognibene et al., 2019; Pushp et al., 2017). While these systems demonstrate proficiency in high-stakes event prediction, they often lack the ability to provide justifications that clarify the motivations behind predicted intentions. The need for a human-like explanation should be considered to elevate system prediction toward a cognitive understanding of why certain outcomes are predicted and how they relate to the broader context of high-stakes situations (Sahoh & Choksuriwong, 2023). It will enhance decision-making by ensuring that AI predictions are effectively understood and appropriately used.

3.2 Human Behavior Explanation

We outline work on social attribution, which defines how people attribute and explain others' behavior. Social attribution focuses not on the actual causes of human behavior but on how individuals attribute or explain the behavior of others. Heider (1958) defines social attribution as person perception, emphasizing the importance of intentions and intentionality. An intention is a mental state where a person commits to a specific action or goal. People consistently agree on classifying events as either "intentional" or "unintentional" (Malle & Knobe, 1997). It is argued that while intentionality can be objective, it is also a social construct, as people ascribe intentions to one another, impacting social interactions.

In addition to intentions, research suggests that other factors, such as beliefs, desires, and traits, play a significant role in attributing social behavior. Researchers from various fields have converged on the insight that people's everyday explanations of behavior are rooted in a basic conceptual framework, commonly referred to as folk psychology or theory of mind (Heider, 1958; Horgan & Woodward, 2013; Malle, 2006b). Folk psychology involves attributing human behavior using everyday terms like beliefs, desires, intentions, emotions, and personality traits. This area of cognitive and social psychology acknowledges that, although these concepts may not genuinely cause human behavior, they are the ones people use to understand and predict each other's actions (Malle, 2006a).

Malle (2006b) presents a model grounded in the Theory of Mind to explain how people attribute behavior to others and themselves by assigning mental states such as desires, beliefs, values, and intentions. This model identifies different modes of behavior explanations and their cognitive processes by distinguishing between intentional and unintentional actions (Figure 1). Intentional behavior is typically explained by reasoning over key mental

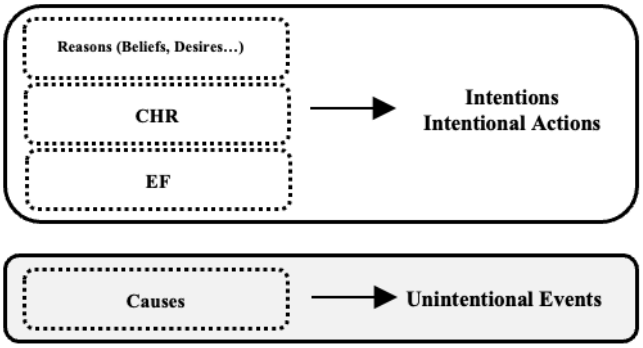


Figure 1: Malle's conceptual framework for behavior explanation; adapted from (De Graaf & Malle, 2017).

components—the reasons behind deliberate acts—based on the rationality principle, where agents are expected to act efficiently to achieve their desires given their beliefs and values. Sometimes, people explain intentional actions using two additional modes: causal history of reason explanations (CHR) and enabling factor explanations (EF). In CHR mode, people focus on factors that influence the reasons behind an action, such as unconscious motives, emotions, and cultural influences, which do not necessarily involve rationality or subjectivity. In EF mode, instead of explaining the intention, they explain how the intention led to the outcome, considering personal abilities or environmental conditions that facilitated the action. Unintentional behavior, on the other hand, is explained by referring to causes such as habitual or physical phenomena, particularly those that prevent or prohibit intentional behavior.

Our proposed model is based on Malle's framework, focusing on the mode of reasoning for intentions and intentional actions. Since people tend to attribute human-like traits to artificial agents, they expect explanations from these agents using the same conceptual framework (De Graaf & Malle, 2017). Therefore, we develop our model by incorporating insights from human-agent studies. To the best of our knowledge, this is the first model designed for explainable GR agents.

4. Human-Agent Study: Insights from Human Explanation

In this section, we present our conceptual model for explaining GR, grounded on empirical data from two different human-agent studies.

4.1 Study Objective

The goal of this study is to investigate how humans explain a GR agent's behavior and identify the key concepts present in such explanations, their frequency, and the relationships among these concepts. Based on our findings, we construct a conceptual framework for GR explanation. We have two case studies with different scenarios and assumptions. The first case study is set in a general domain where no specific expertise is required, such as the Sokoban domain, a classic puzzle game where a player pushes boxes to designated storage

locations within a grid. In contrast, the second case study involves explanations provided by domain experts, as seen in the StarCraft domain, a complex real-time strategy game that requires strategic planning, resource management, and tactical combat.

4.2 Goal Markov Decision Process (Goal MDP)

For both case studies, we employ the Goal Markov Decision Process (Goal MDP) framework to capture an observer's view of the world. A Goal MDP (Ramırez & Geffner, 2011) represents the possible actions that can be taken and the causal relationships of their effects on the world's states. Formally, it is defined as a tuple $\Pi = (S, S_G, A, P, C)$, where S is a non-empty state space, S_G is a non-empty set of goal states, A is a set of actions, $P_a(s' \mid s)$ is the probability of transitioning from state s to state s' given action a, and $C(s, a, s')$ is the cost of that transition. The solution to a Goal MDP is a policy $\pi : S \to A$ that maps states to actions with an overall minimum expected cost.

We use the Goal MDP framework as a cognitive model of the observer's reasoning about goal-directed behavior. In the Sokoban case, the environment is deterministic—each action has a predictable outcome—but we still use the probabilistic transition function to model the observer's subjective uncertainty or confidence in understanding how actions lead to outcomes. In contrast, StarCraft involves actual non-deterministic transitions due to the complexity of real-time interactions, multiple agents, and environmental variability. Here, the probabilistic transition model serves both to capture the true uncertainty in the domain and the observer's perceived uncertainty.

Our approach is based on the following assumptions: (1) the observer perceives the world as a finite set of discrete states and actions; (2) in deterministic domains (like Sokoban), the observer may still interpret transitions probabilistically due to limited knowledge or experience, while in non-deterministic domains (like StarCraft), the uncertainty is inherent; (3) the observer values actions based on costs, aiming to minimize them over time; (4) the observer's reasoning is grounded in their internal world model, which may differ from the actual environment; (5) the observer cannot access the agent's preferences or internal state and must infer intent from observable behavior; and (6) the observer is assumed to have full observability of the environment's current state.

4.3 Case Study 1: Sokoban Game

Sokoban is a classic puzzle game (Figure 2) set in a warehouse environment, where the player or agent navigates through a grid-like layout to move boxes onto designated storage locations. The objective is that each box must be pushed, one at a time, to its assigned spot. The challenge lies in navigating movement constraints and spatial limitations; boxes can only be pushed into empty spaces and cannot be pulled or pushed against walls or other boxes. We modified the Sokoban game rules to allow the player to push multiple boxes simultaneously. This modification transforms the game from a straightforward navigational task into a strategic challenge with multiple objectives, where the player aims to minimize the number of steps taken. We used a STRIPS-like discrete planner to generate plan hypotheses derived from the domain theory and observations as our ground truth.

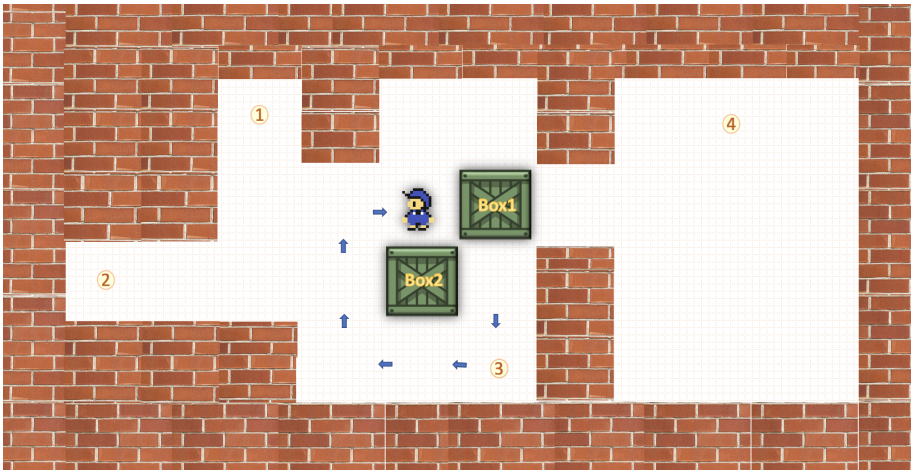


Figure 2: Screenshot of the Sokoban game. There are four storage locations (marked as yellow circles). The possible goals involve delivering two boxes to any combination of these locations. Blue arrows represent observations of the player's move so far.

4.3.1 Study Design

We defined the observer's worldview within the framework of Goal MDPs as follows:

- State Space S: Represents the snapshot of the world, including the player's position, walls, boxes, and storage locations at any given point in time.
- Action Space A: Encompasses the various actions the player can take, either moving or pushing the box(es) in one of the following directions: up, down, right, or left.
- Cost Function $C(s, a, s')$: Assigns a cost for each action to encourage making the fewest amount of moves.
- Goal State S_G: Includes all possible goals that the player can achieve within the current state of the world. Each goal represents a desired configuration of boxes in relation to storage locations.

Table 1 presents different goal representations based on the original game version. In versions 1 and 2, the player can push one box at a time, whereas in version 3, they can push two boxes at a time to goal locations. The goal recognition problems include a number of competing goal hypotheses: multiple possible plans to achieve a goal, two sequential goals with interleaved plans to achieve them, or failed/unsolvable plans (See Appendix A for screenshots of different scenarios).

We additionally varied scenarios to present different *rationality levels*, including rational (optimal or suboptimal) behaviors and irrational behaviors. In total, six observed plans were used in the Sokoban study: three were optimal, two were suboptimal, and one was irrational. For optimal behavior, we assumed that observers would form a simple notion of optimality, where the player follows the least costly plan (i.e., the shortest path in terms of cost) toward a goal. In the suboptimal behavior scenario, the player either chooses a longer path toward

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.

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 included irrational behaviors in which the player fails to complete the task (e.g., getting stuck in a dead-end state) to observe how participants would interpret these actions.

The participant's task was divided into the following phases:

1. Watch an instructional video to introduce the task and game rules.

2. For each of the 18 different scenarios (three games, six scenarios per game):

    - Watch a video clip in which a player tries to achieve the task (Figure 2).
    - After watching the observed actions sequence (plan's completion percentage $min = 25\%$, $median = 53\%$, $max = 83\%$), predict which goal location the player is trying to get to, and, accordingly, assign a likelihood (with one as the least likely and five as the most likely) of each goal. This prediction task is not central to the objectives of this study but was used to engage participants in reasoning about behavior.
    - Provide reasons for your prediction. Participants were required to answer specific questions based on the condition they were in.

Each participant was randomly assigned to one of the following three conditions:

- **Why condition**: Participants were asked: "Explain *why* you have 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 to explain *both why* and *why you have not* in that order.

We collected data for the first and second conditions to analyze the differences between *why* and *why not*, and for 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*.

4.3.2 Data

We recruited 36 participants (22 male, 14 female), allocated evenly and randomly to each condition, aged between 20 and 65, with a mean age of 38. We limited the study to participants from the United States who are fluent in English. Recruitment was conducted via Amazon Mechanical Turk. The participants were compensated $6.50 for completing the task and a bonus of $3.50 for providing more thoughtful answers.

With three different game versions, six scenarios per game, and 12 participants per condition, a total of 864 textual data points were collected (Table 2). We used several methods to filter out deceptive participants. We excluded explanations with fewer than three words or containing gibberish. We also used the time taken to complete the survey as a threshold. This left us with a total of 828 explanations.

We used participants' open-ended explanations to better identify the concepts they used to explain the player's predicted goal. The word count of given answers within the dataset is between 1 and 98 words ($\bar{x}_1 = 22.96$, $\sigma_1 = 15.52$) for the first condition, 3 and 81 words ($\bar{x}_2 = 26.57$, $\sigma_2 = 15.63$) for the second condition, and 1 and 64 words ($\bar{x}_3 = 20.38$, $\sigma_3 = 11.65$) for the third condition.

	#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

4.4 Case Study 2: StarCraft Game

StarCraft is a real-time strategy (RTS) game (Figure 3) where players manage an economy, produce units and buildings, and compete for control of the map with the ultimate aim of defeating all opponents. As an RTS game, StarCraft has several defining characteristics (Ontanón et al., 2013):

- Players engage in a Simultaneous Move Game, where they execute actions concurrently—such as moving units, building structures, and managing resources—demanding effective multitasking skills;
- The Partially Observable Domain limits players' visibility of the game map and opponents' setups, necessitating strategic reconnaissance for informed decisions;
- Real-time gameplay adds urgency, requiring rapid thinking and reflexes to outmaneuver opponents within time constraints;
- Non-Deterministic Actions introduce uncertainty, challenging players to adapt strategies dynamically;
- The game's High Complexity stems from its vast state space and diverse strategic options involving units, buildings, and technologies, compelling players to consider multiple factors when planning strategies.

These elements combine to create a deeply strategic and challenging gaming experience, where success depends on a player's ability to react at strategic, economic, and tactical levels.

Figure 3: Screenshot of the StarCraft game

During gameplay, shoutcasters (commentators) in esports deliver real-time explanations and commentary to audiences, elucidating the intricate strategies and tactics of the game to enhance accessibility and engagement. Their expert insights can potentially benefit XAI tools by analyzing their commentary in the RTS domain (Penney et al., 2021).

4.4.1 Study Design

We defined the worldview of shoutcasters, i.e. observers, within the framework of Goal MDPs as follows:

- State Space S: Encapsulates the game environment's configurations and conditions at any given time. This includes the positions of units, resources, and other relevant game elements.
- Action Space A: Includes low-level actions of specific game units and high-level actions related to game strategies and tactics (sequences of actions).
- Cost Function $C(s, a, s')$: Quantifies the value of states and actions in terms of progress toward winning.
- Goal State S_G: Comprises both subgoals (intermediate objectives that players or teams aim to achieve) and the main goal (the ultimate objective, typically winning the game).

We analyzed a dataset comprising professional StarCraft tournament videos (Penney et al., 2021), where expert shoutcasters provided commentary. We identified key instances where shoutcasters made predictions about players' goals and strategies and coded these instances

to capture the underlying concepts used in their explanations. We clustered the dataset to ensure representative sampling.

4.4.2 Data

We obtained the dataset from Penney et al. (2021), which was collected from professional StarCraft tournaments available as videos on demand from 2016 and 2017. They selected 10 matches and then randomly chose one game from each match. Each of the 10 videos features two shoutcasters (expert commentators) providing commentary.

To obtain representative samples, we identified six clusters within the dataset (1387 instances divided by 6 clusters equals approximately 231 instances per cluster). We then randomly selected one sample of 50 instances from each cluster. The resulting sample size was 300 instances (6 clusters * 50 instances each). We only considered instances involving predictions, specifically when shoutcasters explain their recognition process of what the agents/players aim to achieve (their goals) and so ended up having a total of 132 instances out of the six samples. As the data source is public, we provided supplementary material of coded data to support future research.

4.5 Method

We used a hybrid approach of deductive and inductive reasoning, employing thematic analysis as outlined by Braun and Clarke (2006) to analyze our data. The analysis process was divided into six phases: familiarization with the collected data, development of codes, sorting different codes into potential themes, reviewing themes, defining and naming themes, and writing up the report.

Initially, the collected data was re-read multiple times to ensure immersion before proceeding to the coding phase. The coding process was inductive, aiming to identify basic concepts on how people explain goals, and deductive by relating them to the existing literature on explaining human behavior (Malle, 2006a). Malle's explanation model (Malle, 2006a) shows that people reason over others' beliefs, values, and desires to explain intentional actions. Following that model, we apply these concepts to our coded data. Subsequently, the codes were grouped into defined themes based on their similarities.

In the context of this study, the proposed themes are linked to our research topic of explaining human behavior in goal recognition scenarios. After establishing a set of candidate themes, the refinement process focused on ensuring internal homogeneity—i.e., a cohesive pattern within each candidate theme to accurately reflect the overall data set. Relationships, links, and distinctions between themes were identified during this phase.

The next steps involved naming and describing the themes and illustrating the thematic elements with examples. To ensure consistency, four authors independently coded 10% of the data, achieving approximately 75% inter-rater reliability, as measured by percentage agreement. After achieving this level of agreement, the first and fourth authors continued to code the remainder of the data.

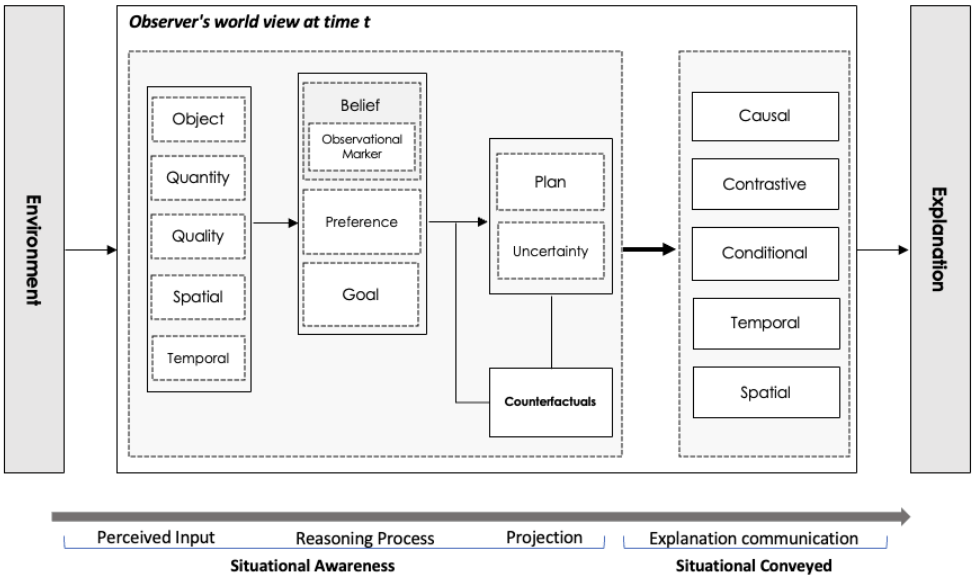


Figure 4: The conceptual model of explainable goal recognition

4.6 Results

4.6.1 The Conceptual Model of Goal Recognition Explanation

We developed the conceptual model by integrating insights from both studies. Initially, we built a model based on the findings from the Sokoban study. This model was then extended to encompass the complexities of the Starcraft study, ensuring that it accurately reflects both the dynamics and unique aspects of each context. The concepts presented in Figure 4 were derived through our thematic analysis process, following the hybrid deductive–inductive approach previously described. Specifically, they emerged from the coded data, where recurring patterns and themes related to observer understanding, explanation types, and contextual factors were identified and refined. The figure highlights the common elements (concepts) that encode an observer's view of the world and the different representations of given explanations.

The model process is guided by two levels; situational awareness and situational conveyance. The situational awareness model, presented by Endsley (1995), is a widely used situation awareness model consisting of three consecutive stages:

1. **Perceived input:** In this stage, the observer perceives the basic elements and their properties in the environment—object, quantity, quality, spatial and temporal information.

2. **Reasoning process:** Based on the perceived inputs, the observer makes causal inferences between their beliefs (including the actor's mental model) and goals, generates counterfactuals, and associates them with their preferences.

3. **Projection:** This stage presents the observer's predictions of future actions guided by their expected goals and the uncertainty level based on the understanding of the previous stages.

In this level, two types of reasoning occur due to the two mental model representations (Felli et al., 2015). The first type is stereotypical reasoning, in which the observer reasons about others' mental states (what the observer would have done). An example from the data corpus is: "It might just do old classic seven gate [a game strategy]." The second type is empathetic reasoning, where the observer casts itself into the actor's mental model and reasons as they would (what the actor would have done). An example is: "This is exactly what I was talking about, you do something to try to force them." In practice, the observer often contrasts both views in a single explanation: What they think should happen vs. what the actor is likely to do. For example: "It's actually going to look for a run by here with this scan it looks like but unfortunately unable to find it with the ravager here poking away." This was observed when explanations were provided for either a failed plan or a sub-optimal one toward achieving a goal. For simplicity, we assume a local perspective in this work, where the actor's mental state is equal to the observer's through the data coding and model implementation process.

The second level is a situational conveyance, where different explanations are formed and communicated by the observer. At this level of the model, the explaining process requires additional strategic knowledge. This includes reasoning about contrastive, conditional, temporal, and spatial cases of problem-solving tasks, allowing for more than just a causal representation of a given explanation.

4.6.2 Concepts

Table 3 shows the different codes and concepts that emerged from the two studies. The given explanations include factual and experiential knowledge ('belief'), subjective likes and objective assessments ('preference'), the desired state to be achieved ('goal'), and possible future actions ('plan'). When people explain others' actions, they infer their goals to provide better explanations (McClure & Hilton, 1997). When explaining a recognized goal, people infer the most relevant evidence from their belief state. Thus, we break down the belief concept into an 'observational marker', an observed precondition that most influences goal prediction. This concept applies not only to optimal behavior, measured by traditional efficiency metrics such as time and shortest route, but also to suboptimal behavior.

Since recognition problems activate counterfactual thinking (Epstude & Roese, 2008), explanations of GR reasoning also include 'counterfactuals' — observational markers, plans, and goals. Explaining counterfactual plans implies having counterfactual goals where the actor has no plan to achieve them. We introduced an uncertainty code as we found that observers use words expressing uncertainty to indicate their confidence level. Additionally, as the problem is to explain goals, 'goal' and 'counterfactual goal' codes were also included. Finally, our data show the presence of different reasoning processes in the explanations. Observers associate a simple causal, conditional, contrastive, temporal, or spatial relationship when generating explanations. A combination of different concepts was used to form the given explanations.

Table 3: Codes (of the concepts) and their descriptions of explanations across two human studies, with examples given from different participants.

Code	Description	Example
Causal	A cause and effect relationship	"**because** it was positioned closer"
Conditional	A conditional relationship	"**if** jjakji takes a 3rd"
Contrastive	A contrastive relationship	"blocked for any goal **but** 1"
Temporal	A series of events over time (order, repetition, opportunity chain, timing)	"a very strong timing that he can hit where he might **be able to** kill"
Spatial	Refer to places or distances	"he has moved **above** it"
Preference/Judgement	Assessment of actions or outcomes	"he gets the **perfect** split"
Goal	Refer to a goal state(s)	"it likely wanted to go to **goal 1**"
Plan	Refer to a future action or sequence of actions	"he is **going to snipe** the warp prism"
Counterfactual goal	Refer to counterfactual goal(s) that could have occurred under different conditions	"boxes have been moved away from **goal one**"
Counterfactual Plan	Refer to a counterfactual action or sequence of actions	"instead of **building a robo in a prism**"
Belief	Refer to general domain knowledge	"**No vision** on the left-hand side of the map"
Observational Marker	Refer to the observed precondition(s) that supports the hypothesized goal(s)	"Given **the player's last move**, box 1 belongs on goal 3"
Counterfactual Observational Marker	Refer to the observed precondition(s) that is against the hypothesized counterfactual goal(s)	"The player **would have taken different steps** if position 1 was the goal"
Object	An object in the domain	"to build a **gate**"
Quantitative	Refer to some quantity or measured value of an object	"a turret at **90%** complete"
Qualitative	Refer to some quality or characteristic of an object	"here with **inferior** roaches"
Uncertainty	Refer to a state of being uncertain	"I **suspect** it is to push the block to goal 2"

Given that we conducted two distinct studies, we developed specific codes tailored to each. In the context of the Sokoban game, the code 'counterfactual observational marker' is explicitly used in the explanations. This code refers to observed preconditions that are against the counterfactual goals. However, observers in the StarCraft game did not incorporate this concept in their explanations, likely due to time constraints and the assumed expertise level of the StarCraft audience.

In the StarCraft game, observers added additional object properties, such as quality and quantity, to describe the characteristics and measured values of objects. This detailed level of description is attributed to the greater complexity of the StarCraft game compared to Sokoban. Furthermore, observers included general domain knowledge (referred to as "belief") in their explanations to clarify facts that were inaccessible to the audience due to the partial observability of the game domain.

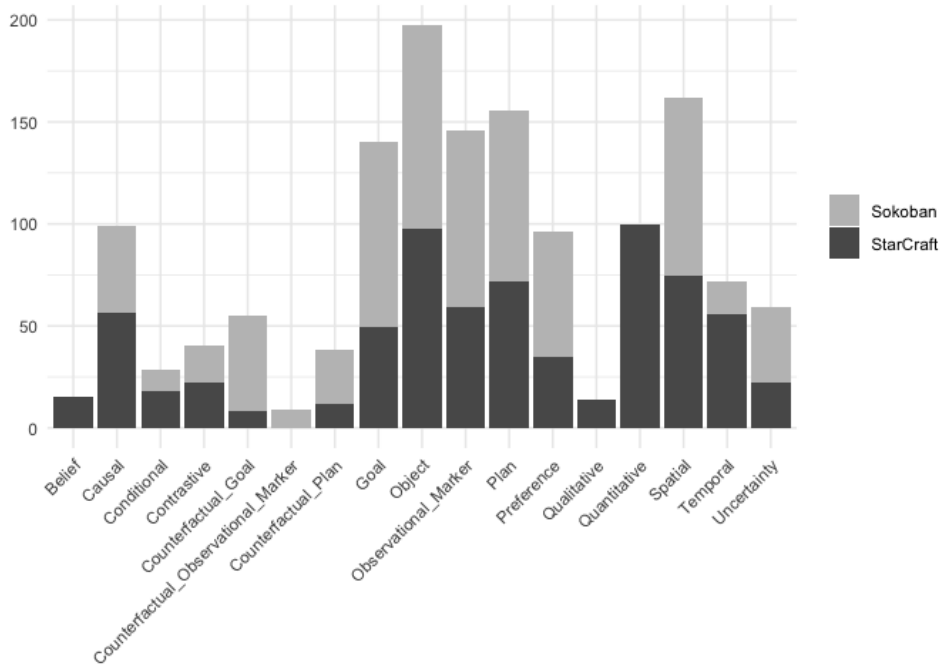


Figure 5: Codes and their frequencies (%) across the two human studies

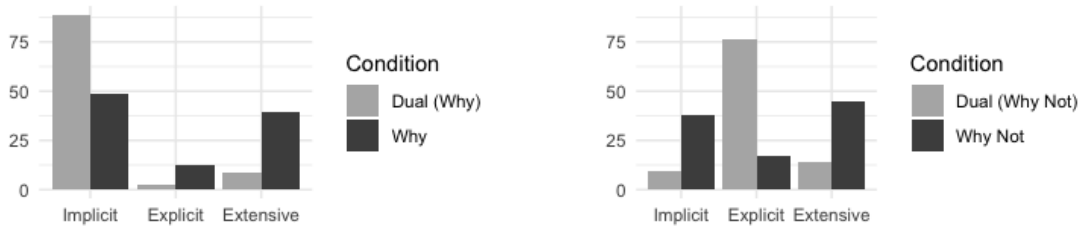
4.6.3 Frequencies

Figure 5 illustrates the frequency distribution of 17 codes across two case studies. Given the navigational nature of both domains, observers predominantly referenced objects and their spatial properties to reflect the players' strategies.

Among the various codes, the observational marker emerges as the most significant finding. This code, which ranks fourth in frequency, provides crucial insights into how observers infer players' intentions and strategies. For instance, in the Sokoban game dataset, an observer noted, "The player positioned itself on top of the box, leading me to believe it is going to push down on the box to reach goal 2." Here, the action "positioned itself on top of the box" is used to explain the entire observed sequence, forming a critical precondition for achieving the predicted goal. This demonstrates how a single observed action can be pivotal in understanding the player's overall strategy. The counterfactual observational marker was coded only in the Sokoban game, as participants tended to use it when responding to 'why not' questions. In the dynamic and time-constrained environment of live StarCraft

commentary, shoutcasters were observed to focus heavily on observational markers in their explanations, even when addressing 'why not' questions.

In addition, the third most frequently occurring code is the 'plan' code, where the actor's goal is explained in terms of how future actions (plans) contribute to achieving that goal. This explanation, provided in terms of future actions, offers insights into potential actions and their execution (Norling, 2009). We believe that people tend to explain in terms of future actions as a way to resolve uncertainty. Causality is also frequently involved, as observers often associate causal relationships to generate explanations.



(a) Answering Why Question

(b) Answering Why Not Question

Figure 6: Explanation Modes and their frequencies (%) over three conditions: Why, Why Not, and Dual.

4.6.4 Questions

In human studies, two forms of causal reasoning are used to answer certain questions (Hoffman & Klein, 2017): retrospective reasoning involves explaining past events through counterfactual reasoning, which considers what could have happened if the observed facts were different, and prospective reasoning involves explaining future events through transfactual reasoning, which considers what could happen in the future if certain conditions are met.

In the Sokoban game, we asked two questions: 'Why?' and 'Why Not', since it has been proved that they are the most demanded explanatory questions (Lim et al., 2009). "Why" questions typically demand contrastive explanations, which are addressed through counterfactual reasoning (Miller, 2019a). In such explanations, people answer 'Why A' in the form of 'Why A instead of B?', where B is some counterfactual goal(s) that did not happen. From the data collected, we classified the provided contrastive explanations into three categories: *implicit*, where observers implicitly contrast and identify relevant causes for A (the predicted goal(s)); *explicit*, where observers explicitly contrast and identify relevant causes for B (the counterfactual goal(s)); and *extensive*, where observers provide explanations for both by identifying relevant causes for A and also for B. It is important to note that the observer answered a "why" question in the 'why' condition, a "why not" question in the 'why not' condition, and both questions sequentially in the 'dual' condition.

Figure 6 illustrates the differences between conditions. In the dual condition, observers tended to adopt an implicit mode when they answered the why question–after having answered why-not—more frequently compared to the why' condition. A similar trend was observed for the why-not question in the dual condition, where observers tended to adopt

an explicit mode when they answered the why-not question–after having answered why —more frequently compared to the Why-not condition. The contrastive nature of these explanations becomes particularly evident in the 'dual' condition, where observers can differentiate between the two types of questions.

Participants primarily engaged in transfactual reasoning when they were highly uncertain about the agent's goal. This uncertainty is a result of the agent's observed behavior being suboptimal (irrational) to all goal hypotheses. For example, from the dataset: "If the player keeps pushing the two boxes together, it would be impossible for box 2 to be put back onto a goal".

Additionally, we coded implicit questions that shoutcasters tried to answer through their predictions in the StarCraft game. There are no pre-specified questions for the observers (shoutcasters) to answer. Instead, they gather information and craft explanations to address the audience's implicit questions. We coded implicit questions that shoutcasters tried to answer through their predictions. Table 7 shows these questions and their frequencies. The shoutcasters are primarily focused on reasoning prospectively, addressing the "What could happen?" question allows them to anticipate future events. Despite the time constraints, they managed to provide reasoning for their predictions by answering other questions of interest.

The most frequently answered question is the "Why?" question, involving retrospective reasoning over past events that mostly influence goal prediction. They further support their predictions by anticipating factors that control the context, allowing them to prospectively answer the "How?" question through future projections.

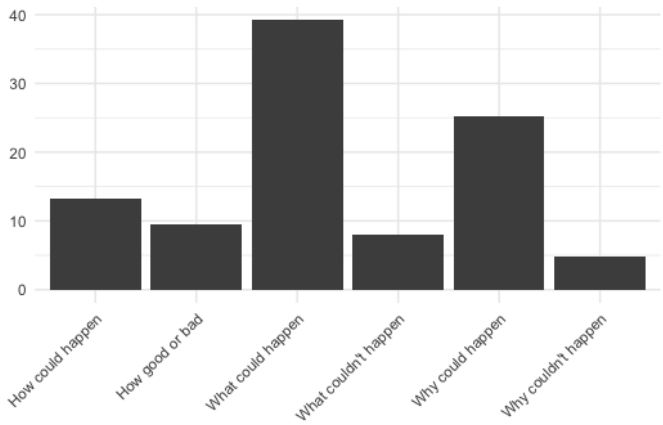


Figure 7: Questions and their frequencies (%) in StarCraft game

4.6.5 Discussion

Causal reasoning is essential for constructing mental representations of events (Luo & Baillargeon, 2010; Malle, 2006a; Pearl et al., 2000). These representations form causal chains that illustrate how a sequence of causes leads to an outcome. In the realm of explainable AI, an agent aiming to explain observed events may need to use abductive reasoning to

identify a plausible set of causes (Miller, 2019b). While numerous causes can contribute to an event, individuals typically select a subset they deem most relevant for their explanation (Miller, 2019b).

Our human studies focus on understanding how people explain others' predicted goals based on observed behavior. By coding these explanations, we identify a key concept, referred to as an 'observational marker' that participants use to build their explanations. Our findings align with social and cognitive research indicating that people prefer explanatory causes that seem sufficient in the given context for the event to occur (Lipton, 1990; Lombrozo, 2010; Spellman, 1997; Woodward, 2006).

People make causal inferences about others' beliefs and goals based on their observed behavior and prior domain knowledge (Baker et al., 2009). A key aspect to explain those inferences is the ability to decide to what degree the observed evidence from a causal chain supports a goal hypothesis. To this end, we propose an explanation model for GR agents based on concepts such as causality and observational markers.

5. eXplainable Goal Recognition (XGR) Model

Building on insights from human studies discussed in Section 4 and the conceptual model presented in Figure 4, we propose a simple and elegant explainability model for goal recognition algorithms called *eXplainable Goal Recognition* (XGR). The XGR model formalizes two core components identified in the conceptual model—observational markers and counterfactuals—using the Weight of Evidence (WoE) framework. Observational markers reflect the observed precondition(s) that supports the hypothesized goal, while counterfactuals support reasoning about alternatives. Together, they allow the XGR model to produce interpretable structured explanations for goal hypotheses. In the following section, we use the navigational GR example (Example 5.1) as a running example to support the definitions.

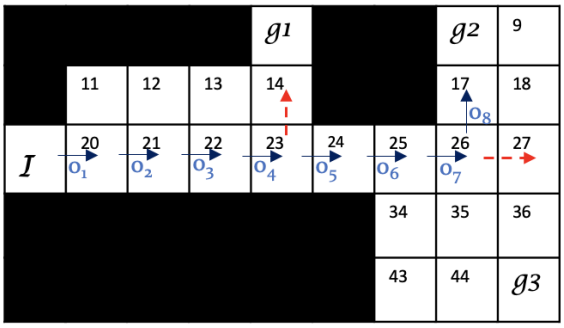


Figure 8: Navigational domain example.

5.1 Running Example

Figure 8 presents a navigational domain where an agent can navigate through the unblocked grid to reach one of 3 possible goal locations. The GR task is composed of an initial state where an agent at the start is located (marked I), a set of goal hypotheses, $G = \{g_1, g_2, g_3\}$, and a sequence of observations $O = \langle o_1, .., o_8 \rangle$ (represented as blue arrows). The domain

definition $\Xi = \langle F, A \rangle$ includes a fact set F comprising the cells (45 states in total) and an action set A defined by four types of moves: up, down, left, and right, all with equal cost. The domain model is deterministic and discrete, meaning each action has only one possible outcome — although our model does not assume deterministic actions. A goal state specification G is defined as the agent being in one of the three possible goal cells. Thus, the Mirroring GR (see Section 2.2.1) would infer that, most likely the agent's goal is to reach g_2 since the observed sequence confirms the optimal plan to achieve this goal.

5.2 Overview

Extending Melis et al. (2021) WoE framework, our model addresses the 'why' and 'why not' questions, which are the most demanded explanatory questions (Lim et al., 2009). Lim et al. (2009) compared a range of intelligibility-type questions and showed that explanations describing *why* a system behaved in a certain way resulted in better understanding and increased trust in the system, which is also supported by our findings in Section 4.6.4 (Figure 7).

The model accepts four components as input, which any GR model can provide:

1. An observed sequence O_i, representing the set of observations observed up to and including the current time step i.
2. An observation $o_i \in O_i$, which is the most recent piece of observation being considered.
3. The set of predicted goals, G_p, where $G_p \subseteq G$.
4. The set of counterfactual (not predicted) goals, G_c, where $G_c \subset G$, where $G_p \cap G_c = \varnothing$ and $G_p \cup G_c = G$.
5. Posterior probabilities $P(g \mid O_i)$ for each goal $g \in G$ and $o_i \in O_i$, indicating how likely each goal is given the observation up to and including o_i.

We define an explanation as a pair that contains: (1) an explanandum, the event to be explained (that is, goal hypotheses); and (2) an explanan, a list of causes provided as the explanation (Miller, 2019a). The model answers two questions: 'Why goal g?', where g is a predicted goal hypothesis, and 'Why not goal g'?', where g' is a counterfactual goal hypothesis. Assuming a full observation sequence, the explanation is given as a list of an observed action and its WoE value. Given that list, the selection of the explanations is based on the type of question to be answered.

5.3 Explanation Generation

We generate explanations by extending the WoE framework presented by Melis et al. (2021). By generating explanation lists using WoE, we can measure the relative importance of each observation to the goal hypotheses. This approach enables us to explain using the *observational marker*, which is the predominant concept used in the explanation of the agent's goals (refer to Section 4.6.3).

Referring to Equation 1, we substitute the hypotheses h and h' with a predicted goal and counterfactual goal hypotheses, g and g'. The evidence e is replaced by an observed

Algorithm 1 Explanation Generation Algorithm

Input: O_i, o_i, G_p, G_c, and Posterior probability over G

Output: Explanation list Ω for all pairs (G_p, G_c)

```
Ω ← [] {Initialize explanation list}
for o_i ∈ O do
    for g ∈ G_p do
        for g' ∈ G_c do
            ω_i ← woe(g/g' : o_i | O_i)) {Compute Weight of Evidence (WoE)}
            Ω ← Ω ∪ {(g, g') = ⟨ω_i, o_i⟩} {Add explanation to list}
        end for
    end for
end for
return Ω
```

action o_i and the posterior probabilities are represented as $P(g \mid O_i)$ and $P(g' \mid O_i)$. A complete explanation is defined as follows:

Definition 5.1. A *complete explanan* for a goal g is a list of pairs $(woe(g/g' : o_i \mid O_i), o_i)$, in which the conditional $woe(g/g' : o_i \mid O_i)$ for each paired hypothesis g and g' is computed for each added observation o_i to the observed sequence O_i. The WoE is computed as follows:

$$woe(g/g' : o_i \mid O_i) = \log \frac{P(g \mid O_i)}{P(g' \mid O_i)} \tag{2}$$

Informally, this defines a complete explanan for a goal g as the complete list of computed WoE scores for each observation. An algorithm for extracting this is shown below (Algorithm 1).

In the navigational GR scenario presented previously (Figure 8), the WoE would be the same for all goal hypotheses after the first three observations, o_0 to o_3. This is because the Mirroring GR algorithm predicts them as equally likely since the first three observations are part of the optimal plan to achieve all three goals. However, this uniformity does not hold for the rest of the observation sequence. For observations o_4 to o_6, the mirroring GR outputs would be goal $g2$ and $g3$ as equally likely since the observed actions are consistent with the optimal actions needed to reach either goal, with the counterfactual goal being $g1$. Table 4 presents the posterior probabilities and WoE values associated with either g_2 or g_3 as the leading goal candidate. The model computes the WoE value of each observed action for the pair of the predicted and counterfactual goals.

5.4 Explanation Selection

Explaining the output of a GR algorithm in terms of the *complete* observation sequence can be tedious or even impossible, especially in scenarios where the domain model contains hundreds of thousands of states and actions. XAI best practice deems that for explanations to be effective, they should be selective, focusing on one or two possible causes instead of all possible causes for a decision or recommendation (Miller, 2019a). In the context of GR explanations, we found that people pointed to the *observational marker* and the *counterfactual observational marker* when they answered 'why' and 'why not' questions

(refer to Sections 4.6.2, and 4.6.3). To this end, we focus on selecting explanations to answer 'Why g?' and 'Why not g'?' questions.

$o_i \in O$	g	g'	$P(g \mid o_i)$	$P(g' \mid o_i)$	$woe(g/g' : o_i)$
o_4	g_2	g_1	0.36	0.27	0.28
	g_3	g_1	0.36	0.27	0.28
o_5	g_2	g_1	0.38	0.23	0.51
	g_3	g_1	0.38	0.23	0.51
o_6	g_2	g_1	0.40	0.20	0.69
	g_3	g_1	0.40	0.20	0.69

Table 4: Posterior Probabilities and Weight of Evidence (WoE) for Predicted and Counterfactual Goals After Observations o_4, o_5 and o_6 in the Navigational GR Example Depicted in Figure 8

5.4.1 'Why' Questions

Answering *why goal g*? questions, such as *Why was goal g predicted as the most likely goal over all other alternatives*?, rely on identifying the most important observation(s) that support the achievement of that goal. We call such observations *observational markers* (OMs).

Definition 5.2 (Observational Marker)**.** Given a complete explanan of g, the *observational markers* (*OMs*) are the observed actions that have the highest WoE value:

$$OM = \arg\max_{o_i \in O_i} [(g, g') = \langle \omega_i, o_i \rangle] \tag{3}$$

It is generated for every possible alternative, and in case of having multiple such actions, we select them all. Consider the navigational GR scenario presented in Figure 8. Let us answer the question *Why* g_2?. From the *complete explanan* of g_2, shown in Table 4:

$$\begin{aligned}(g_2, g_1) = &\ [\langle 0.28, o_5 \rangle, \langle 0.51, o_6 \rangle, \langle 0.69, o_7 \rangle, \langle 0.85, o_8 \rangle] \\ (g_2, g_3) = &\ [\langle 0.18, o_8 \rangle]\end{aligned}$$

After ranking them from highest to lowest, we obtain $\langle 0.85, o_8 \rangle$ that has the highest value. This indicates that this observation is the OM, as in the observation that best explains the predicted goal hypothesis $G_p = \{g_2\}$ instead of the counterfactual goal hypotheses, $G_c = \{g_1, g_3\}$. Therefore, the explanation would be *Because the agent has moved up from cell 26 to cell 17.*

5.4.2 'Why Not' Questions

The question of *why not* g'? relies on identifying the most important observation(s) related to g', which are called *counterfactual observational markers.*

Definition 5.3. Given a complete explanan of g', the *counterfactual observational markers* (*counterfactual OMs*) are the observation(s) that have the lowest WoE value:

$$\text{counterfactualOM} = \arg\min_{o_i \in O_i} [(g, g') = \langle \omega_i, o_i \rangle] \tag{4}$$

There may be multiple such observations, in which case we select all of them. Consider the navigational GR scenario (Figure 8) and the question *Why not* g_1 *and* g_3? From the *complete explanation* of g_1 and g_3, shown in Table 4:

$$\begin{aligned}(g_2, g_1) = &\ [\langle 0.28, o_5\rangle, \langle 0.51, o_6\rangle, \langle 0.69, o_7\rangle, \langle 0.85, o_8\rangle]\\ (g_2, g_3) = &\ [\langle 0.18, o_8\rangle]\end{aligned}$$

After ranking them from lowest to highest, we obtain $\langle 0.28, o_5\rangle$ as the lowest value for g_1 and $\langle 0.18, o_8\rangle$ as the lowest value for g_3. This indicates that these observations are the *counterfactualOM*, the observations that best explain the counterfactual goal hypotheses, $G_c = g_1, g_3$. Therefore, the explanation would be: *Because the agent moved right from cell 23 to cell 24, away from* g_1*, and up from cell 26 to cell 17, away from* g_3.

Counterfactual Action Pointing to the lowest WoE action is not enough to answer *why not* g'. Part of answering 'why not' questions is the ability to reason about the *counterfactual plan* that should have occurred instead of *counterfactual OM* for g' to be the predicted goal (see Section 4.6.2).

Building on this idea, we obtain the counterfactual action that should have happened instead of the observed action by planning the agent's route to g' and simply taking the first action. We approach this problem by generating a plan for $g1$ from the state that precedes the obtaining of the *counterfactual OM*, the state from which the lowest WoE is measured. We define the counterfactual action as follows.

Definition 5.4. Given a *counterfactual OM* o_i at state s_{t-1} for counterfactual goal g', a *counterfactual action* a'_t is any action that appears as the first step in a valid plan $\pi = \langle a'_t, a'_{t+1}, .., g'\rangle$ that solves the planning problem $\langle \Xi, s_{t-1}, g'\rangle$, where Ξ is the planning domain. When multiple valid plans exist, a'_t may be selected from the set of first actions of all such plans.

Consider again the example in Figure 8. The counterfactual action to g_1 would be the *move up* action from cell 23 to 14, and to g_3 would be the *move right* action from cell 26 to 27 (as indicated by the red arrows). Verbally, the answer to "Why not goal g_1?" would be: *Because the agent moved right from cell 23 to cell 24. It would have moved up from cell 23 to 14 if the goal was* g_1, and to "Why not goal g_3?" would be: *Because the agent moved up from cell 26 to cell 17. It would have moved right from cell 26 to 27 if the goal was* g_3

As noted in the navigational example (Figure 8), the counterfactual action for g_3 which is moving right from 26 to 27 (represented as a red arrow) is part of a suboptimal plan to g_2 (moving right to 27, up to 18, up to 9, and left to g_2). The framework operates by identifying the lowest observed evidence for a goal g against goal g' at time step t, and generating the counterfactual action from that point towards g', even if that action is part of a plan for g.

6. Evaluation

In this section, we present a comprehensive analysis of the XGR framework as obtained through a combination of a computational study and three user studies to assess the effectiveness of our model.

6.1 Computational Evaluation

We evaluate the computational cost of the XGR model over eight online GR benchmark domains (Vered et al., 2018) to determine whether the cost of our approach is suitable for real-time explainability. The benchmark domains vary in levels of complexity and size, including different numbers of observations and goal hypotheses. We measure the overall time taken to run the XGR model. As the explanation model uses an off-the-shelf planner for counterfactual planning, we also separate the planner's cost and the explanation generation and show its effect on overall model performance.

Table 5 presents the run time performance of the XGR model over the benchmark domains. The run times vary greatly depending on the complexity of the domain, ranging from an average of 0.14 seconds over the 15 problems in the relatively simple Kitchen domain, to 221.77 seconds over the 16 problems in the complex Zeno-Travel domain (column 1). Regardless of the run time, adding our explainability model to the GR approach is typically not expensive, adding an increase of between 0.2%-45% (column 3). However, most of this increase can be attributed to calling the planner to generate counterfactual explanations (column 4). We can see that between 70%-99% of the XGR model run time is spent on planning. The varying percentage increases between domains like Zeno-Travel and Kitchen emphasize the relationship between domain complexity and planning time: the higher the domain complexity, the greater the influence the planner has. This highlights the significant impact of planner selection on the model performance. Since our model is independent of the underlying GR model, it has the potential to scale effectively with the integration of more efficient planners, such as domain-specific planners.

6.2 Empirical Evaluation

We consider human studies experiments essential to the XGR model evaluation and conduct three human studies. The studies were conducted after obtaining institutional HREC approval.

6.3 Study 1 - Generating Human-Like Explanations

Our first study evaluates whether the model output is grounded on human-like explanations.

6.3.1 Methodology and Experiment Design

We used the method of annotator agreement and ground truth, where human annotations of representative features provided the ground truth for the quantitative evaluation of explanation quality (Mohseni et al., 2020).

Task Setup In this study, participants interacted with the output of the Mirroring GR algorithm across a series of problems within the Sokoban game domain (refer to Section 4.3 for additional details). The game was divided into three versions:

- Game Version 1: Required the delivery of a single box to a single destination.
- Game Versions 2 and 3: Involved delivering two boxes to two sequential destinations, with interleaved plans to achieve each goal. The key distinction between these versions

Domain (# problems)	Mirroring with XGR (sec)	XGR only (sec)	Time Increase (%)	Counterfactual Planning (%)
Campus (15)	0.21 (0.08)	0.019 (0.017)	10.15	87.11
Ferry (24)	71.22 (36.16)	6.276 (8.070)	09.66	99.69
Intrusion (45)	0.69 (0.36)	0.215 (0.087)	44.61	70.18
Kitchen (15)	0.14 (0.07)	0.014 (0.002)	11.12	73.61
Rovers (20)	135.23 (73.11)	3.710 (7.271)	02.82	99.64
Satellite (27)	16.76 (10.05)	1.794 (10.052)	11.98	99.27
Miconic (20)	109.12 (22.61)	1.636 (2.861)	01.52	98.72
Zeno-Travel (16)	221.77 (68.85)	8.856 (11.721)	04.15	99.65

Table 5: Performance results of the XGR model for the *Mirroring* Goal Recognition algorithm across eight benchmark domains. Column 1: Average and standard deviation of runtime with XGR and the mirroring GR algorithm. Column 2: Average and standard deviation of runtime with XGR only. Column 3: Percentage increase in runtime with XGR added to the GR approach. Column 4: Percentage of this additional runtime spent on counterfactual planning.

was that, in version 3, the agent could push multiple boxes, whereas in version 2, the agent could only push one box at a time.

This progression shifted the task from a straightforward navigation challenge to a more strategic one, where the player needed to manage multiple goals while striving to minimize the number of steps taken. Each game version included five scenarios of varying complexity, for a total of 15 scenarios. Each scenario presented a different goal recognition problem, with multiple competing goal hypotheses. Participants were required to answer "why" and "why not" questions regarding the predicted and counterfactual goal sets.

Procedure Each experiment lasted approximately 60 minutes and included the following four stages:

1. The game instructions were introduced to the annotators, along with a training scenario to help them understand the task.

2. The annotators watched a partial scenario (video clip) in which a Sokoban player attempted to achieve a goal (see Figure 9). The goals involved either delivering/pushing a box to a single destination cell or delivering/pushing two boxes to two different destination cells.

Figure 9: Sokoban game, scenario 5 (Game 3). The blue dot marks the initial state of the agent, and the orange dot marks the initial state of the box1 before pushing it down. There are 3 possible goals $((g_1, g_2), (g_3, g_4), (g_5, g_6))$. Blue arrows represent observations of the agent's actions.

3. After watching the incomplete observation sequence, annotators were given the set of predicted and counterfactual goals generated by the mirroring goal recognition algorithm.

4. The annotators were asked to identify the most important action, or optionally the two most important actions, from the observation sequence that addressed the questions: 'Why goal g?' and 'Why not goal g'?' Here, g was the predicted goal, and g' was the counterfactual goal. Additionally, participants were asked to annotate a counterfactual action for 'Why not goal g?'. This involved proposing a *non-observed* action that they believed would indicate a move towards the alternative goal g' (see Appendix C1. for an example scenario screenshot).

Participants We recruited three annotators (one male, two female) from the graduate student cohort at our university. Participants were aged between 29 and 40, with a mean age of 33. No prior knowledge of the task was required.

Metrics Using a majority vote, we combined participants' annotations into a single ground truth. Disagreements between annotations were typically resolved by an extra annotator. Then we calculated the mean absolute error (MAE) to assess how closely the obtained explanation of the XGR model matched the ground truth, defined by the agreement between the annotators. The MAE was calculated as the average difference between each ground truth value ($a_{groundTruth}$) and the corresponding XGR value (a_{XGR}) over the length of the observation sequence (n). These values will be discussed in the following section.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \mid a_{groundTruth} - a_{XGR} \mid \tag{5}$$

o_i	Why Question	WhyNot Question
o_8	1	6
o_7	2	5
o_6	3	4
o_5	4	4
o_4	5	3
o_3	6	2
o_2	7	1
o_1	0	0

Table 6: An explanation list generated by the XGR model which ranks each observation to answer 'Why' and 'Why not' questions for the example scenario depicted in Figure 9.

To evaluate the selection of a counterfactual action, we compute the binary agreement between the XGR model and the ground truth. For each instance, the predicted action is considered correct (assigned 1) if it matches the first action of any valid plan that solves $\langle \Xi, s_{t-1}, g' \rangle$; otherwise, it is marked as incorrect (assigned 0). The overall agreement percentage is then calculated using the method described by Araujo and Born (1985):

$$CF(\%) = \frac{agreements}{agreements + disagreements} \times 100 \tag{6}$$

6.3.2 Results

We applied our XGR model to the online Mirroring implementation for each of the 15 scenarios. To determine the value of (a_{XGR}), we obtained the ranked explanation list from the XGR model based on the Weight of Evidence (WoE) values. For the question 'Why goal g?', we ranked the explanations in descending order of their WoE values, assigning a rank of 1 to the explanation with the highest WoE (Observational Marker, OM). Similarly, for the question "Why not goal g'?", we ranked the explanations in ascending order of their WoE values, assigning a rank of 1 to the explanation with the lowest WoE (counterfactual OM). Using ground truth obtained through human annotation, we matched these ranks with their equivalents in the ranked explanation list for each question to determine $a_{groundTruth}$. We then calculated the Mean Absolute Error (MAE) for each question and across all 15 scenarios.

Example 6.1. Consider the example in Figure 9. We obtained the ranked explanation list for both questions from our model (Table 6). The annotated actions from the ground truth are o_2 and o_7, which explain 'Why g_1 AND g_2?', and o_2, which explains 'Why not (g_3 AND g_4) OR (g_5 AND g_6)?'. We then determined ($a_{groundTruth}$), which is the equivalent rank of the annotated action in the list. For the "Why" question, these values are $o_2 = 7$ and $o_7 = 2$, and for the "Why not" question, the value is $o_2 = 1$.

The Mean Absolute Error (MAE) is the average of the errors; hence, the larger the number, the larger the error. An error of 0 indicates full agreement between the models.

The results of the comparison are presented in Table 7. Each row shows the MAE calculated for each game scenario. The 'Why' and 'Why not' columns represent the MAE

for our model compared to the human ground truth, while the CF(%) column represents the percentage of counterfactual action explanations that agree with the human ground truth.

For the majority of instances, the XGR model agreed with the ground truth obtained through human annotation. When answering 'Why g?' questions, the model had a full agreement with the ground truth in 11 out of the 15 scenarios (73.3%). For 'Why not g'?' questions, the model had a full agreement with the ground truth in 14 out of the 15 scenarios (93.3%). By 'full agreement', we mean that the human annotators identify the same two actions as the primary explanation.

The CF column represents the percentage of counterfactual action explanations that agree with the human ground truth. Higher values are better, with 100% indicating full agreement with the ground truth counterfactual actions. The model achieved full agreement in 11 out of the 15 scenarios (73.3%), demonstrating excellent performance.

Game	**Scenario**	**Why**	**Why Not**	**CF** (%)
1	S1	0.00	0.00	100
	S2	0.00	0.00	100
	S3	**0.37**	0.00	100
	S5	**0.25**	0.00	100
	S5	0.00	0.00	100
2	S1	0.00	0.00	66.6
	S2	0.00	**0.12**	33.3
	S3	0.00	0.00	100
	S4	0.00	0.00	100
	S5	0.00	0.00	33.3
3	S1	**0.50**	0.00	100
	S2	0.00	0.00	50
	S3	0.00	0.00	100
	S4	0.00	0.00	100
	S5	**0.44**	0.00	100
Mean		0.10	0.008	89.40
SD		0.65	0.031	00.25

Table 7: The *Why* and *Why not* columns represent the mean absolute error (MAE) for XGR compared to the ground truth. The CF column represents the counterfactual action explanations percentage that agreed with the ground truth.

To better understand the performance of our model, we investigated scenario 5 in game 3, which has a relatively high Mean Absolute Error (MAE) for the *Why goal g?* question, we refer to Figure 9. In this scenario, the agent delivers two boxes to two different locations and can push two boxes simultaneously. The blue arrows in the figure represent the observation sequence, indicating that the agent started in the blue circle and followed the arrows to its current location.

In this scenario, the most likely goal candidate, as predicted by the Mirroring GR algorithm, was delivering Box1 to g_1 and Box2 to g_2, i.e., $G_p = (g_1, g_2)$. The counterfactual goal candidates involved delivering the boxes to either g_3 and g_4 or g_5 and g_6, with $G_c = (g_3, g_4), (g_5, g_6)$. It is important to note that to push both boxes simultaneously to goal

(g_1, g_2), the agent would need to stand to the right of the boxes. Conversely, to push both boxes simultaneously to either (g_3, g_4) or (g_5, g_6), the agent would need to position itself to the left of the boxes.

The XGR model's explanation for the *Why goal g?* question is observation o_8 in the figure. This observation suggests that the agent aims to position itself to the right of both boxes, confirming the hypothesis that the goal is (g_1, g_2). Considering the agent's ability to push multiple boxes, this observation constitutes the Observation Model (OM) with the highest Weight of Evidence (WoE).

On the other hand, the annotators established the ground truth explanation by choosing the second observation, o_2, for both the *Why goal g*? and *Why not g′*? questions. According to our model, this observation is the one with the lowest Weight of Evidence (WoE), actually making it the *counterfactual observation* and the answer to the question *Why not g′*? This is because this observation moves away from both goals (g_3, g_4) and (g_5, g_6).

Participants choosing to use the same answer for both *Why goal g*? and *Why not g′*? questions can also be found in other instances of discrepancies between the output of our model and the ground truth. The difference in explanations can be attributed to some confusion and/or preference between *why*? and *why not*? questions on the part of the participants.

To address this, we conducted a follow-up experiment where we presented the participants with the scenarios they were confused with (Table 7, scenarios with bold values). For each scenario, we provided them with explanations from two systems: the first system's explanation from our model, and the second system's explanation from the ground truth. We then asked them which system provided a better explanation. All three participants preferred the explanations provided by our model which leads us to our model's explainability as perceived by users.

6.4 Study 2 - Perceived Explainability

The second human subject experiment aimed to evaluate the explainability of our model. We considered the following two hypotheses for our evaluation: 1) Our model (XGR) leads to a better understanding of a GR agent; and 2) A better understanding of a GR agent fosters user trust.

6.4.1 Experiment Design and Methodology

Conditions We conducted a between-subjects study in which participants were randomly assigned to one of two conditions: 1) the explanation model (XGR), where an explanation was provided for the GR output; 2) the No Explanation model, where no explanation was provided for the GR output. We did not include a baseline for another explanation method due to the lack of existing alternatives.

Task Setup We used the Sokoban game as our test bed, following the same task setup as in our previous study (see Section 6.3). The participants engaged with the output of the mirroring GR algorithm in a series of problems within the Sokoban game domain. For each problem, we used a STRIPS-like discrete planner to generate ground truth plan hypotheses based on the domain theory and observations. The participants were then tasked with predicting the player's possible goal based on the observed behavior. Following this, they

rated their trust on a 5-point Likert scale (Hoffman et al., 2018). In the XGR condition, participants also used a 5-point Likert scale to rate the given explanation according to their satisfaction with it (Hoffman et al., 2018).

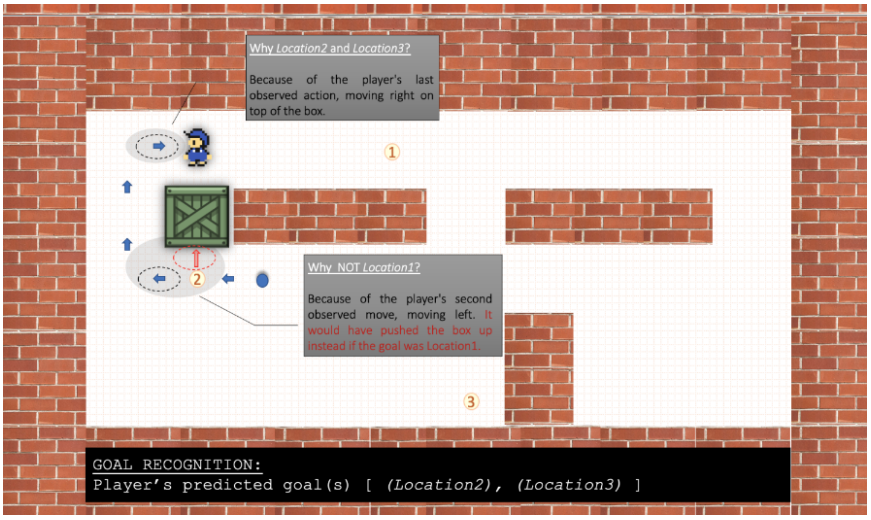


Figure 10: Example scenario of Sokoban game, XGR condition. The blue dot marks the initial state of the agent. There are 3 possible goals ($g1, g2, g3$). Blue arrows represent observations of the agent's actions, and red arrow represents the counterfactual action.

Procedure Participants were presented with six partial scenarios (video clips) showing a Sokoban player attempting to deliver boxes to designated locations. The experiment was divided into four phases:

1. **Phase 1:** Collection of demographic information and participant training. Using two video clips, the participant is trained to understand the player task and how to use GR and explainable system outputs.

2. **Phase 2:** Presentation of a 10-second video clip of the Sokoban player's actions, along with the GR system output. Participants were asked to predict the agent's goals. For the No Explanation condition, participants made predictions without receiving any explanations. In the XGR condition, explanations for 'why' and 'why not' questions were presented (see Appendix C2. for example scenarios of the two conditions). Explanations were pre-generated by our algorithm and displayed on an annotated image of the video clip's final frame. The explanations were converted into natural language using a template, as exemplified in Figure 10.

3. **Phase 3:** Completion of the trust scale by participants.

4. **Phase 4 (XGR condition only):** Completion of the explanation satisfaction scale.

Participants Prior to running the study, we performed a power analysis to determine the needed sample size. We calculated Cohen's F and obtained an effect size of 0.35. Using

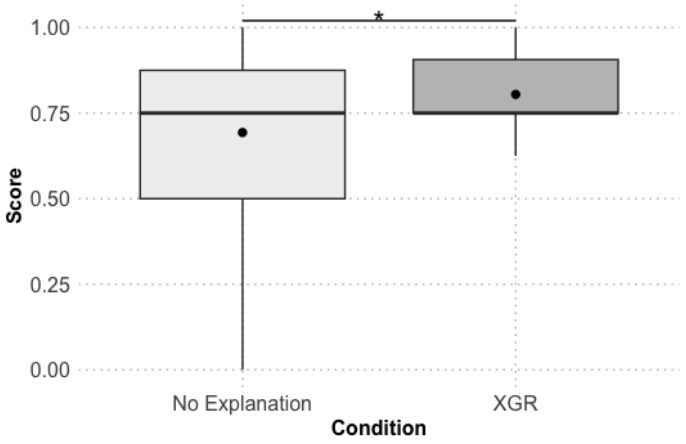


Figure 11: Task prediction scores for the two models (higher is better).

a power of 0.80 and a significance alpha of 0.05, this resulted in a total sample size of 60 for the two conditions. We therefore recruited a total of 70 participants from Amazon MTurk, allocated randomly and evenly to each condition. To ensure data quality, we recruited only 'master class' workers whose first language is English and who have at least a 98% approval rate on previous submissions. After excluding inattentive participants, we obtained 65 valid responses (No Explanation: 33, XGR: 32). Demographics included 28 males and 37 females, aged between 20 and 69, with a mean age of 40. Participants were compensated $4.00 USD, with an additional bonus of $0.20 USD for each correct prediction, up to a total of $1.20 USD.

Metrics To test Hypothesis 1 (XGR leads to a better understanding of a GR agent), we used the task prediction method as described by Hoffman et al. (2018), which acts as a proxy measure for user understanding. Participants were scored one point for each correct prediction and penalized one point for each incorrect prediction.

To test Hypothesis 2 (A better understanding of a GR agent fosters user trust), we used the *trust scale* from Hoffman et al. (2018), where participants rated their trust on a 5-point Likert scale ranging from 0 (Strongly Disagree) to 100 (Strongly Agree) across four dimensions.

Additionally, to evaluate the subjective quality of the explanations, participants completed the *Explanation Satisfaction Scale* from Hoffman et al. (2018), also measured on a 5-point Likert scale from 0 to 100 across four metrics.

Analysis Method After conducting a homogeneity test to assess the variance of the collected data, we proceeded with Welch's t-test.

6.4.2 Results

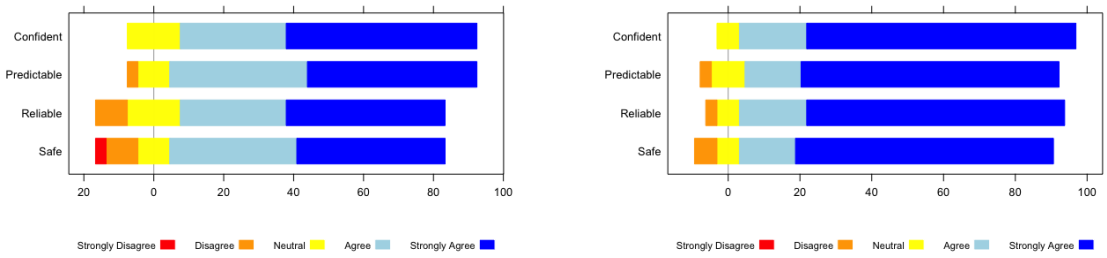
Hypothesis 1: Our model (XGR) leads to a better understanding of a GR agent Figure (11) presents the variance in task prediction scores for the two models. A Welch's t-test indicated a significant difference (p-value = 0.03) in favor of the XGR cohort and the

No Explanation cohort. These results suggest that our model (XGR) provides a significantly better understanding of the agent's behavior compared to the baseline model, as evidenced by the task prediction scores. Therefore, we accept our first hypothesis.

Table 8 shows the mean and standard deviation of the explanation quality metrics for the XGR model on a Likert scale. Higher values indicate stronger agreement. These results suggest a satisfactory level across all four metrics.

Understanding	Satisfying	Sufficient Detail	Complete
88.93 (10.8)	86.09 (12.2)	87.93 (13.1)	86.15 (19.6)

Table 8: Mean and standard deviation of explanation quality metrics for the XGR model



(a) No Explanation Model

(b) XGR Model

Figure 12: Likert scale count of perceived trust metrics for the two conditions. The X-axis represents each Likert category's total counts of responses, adjusted to represent 0 as the midpoint.

Hypothesis 2: A better understanding of a GR agent fosters user trust Figure (12) illustrates the Likert scale data distribution of the perceived level of trust of the participant between the two models. A Welch t-test yielded p-values $(0.10, 0.18, 0.03, 0.02)$ for the trust metrics (*Confident*, *Predictable*, *Reliable*, and *Safe*) respectively. These results indicate a significant difference in the *Reliable* and *Safe*, and a marginally significant difference for the *Confident* metric. These results support our second hypothesis, demonstrating that better understanding of the GR agent, as facilitated by the XGR model, fosters increased user trust.

6.5 Study 3 - Effectiveness in Supporting Decision-Making

This study aims to evaluate our model within the context of human-AI decision-making, focusing on several key hypotheses for empirical assessment. Specifically, we intend to investigate the impact of incorporating counterfactual explanations into our model. Our hypotheses are as follows: 1) Decision Accuracy: We hypothesize that our model enhances decision-making performance; 2) Decision Efficiency: We anticipate that the model will contribute to greater overall efficiency in task completion; 3) Appropriate Reliance on the GR Model: We propose that our model will promote more *appropriate* reliance on the GR model; and 4) User Trust: We expect that our model will lead to increased trust in the

GR agent. For this hypothesis, the aim is to further validate our findings (refer to Section 6.4.2) in a different context. Additionally, we seek to address the research question: *How do participants' reasoning processes vary across the four conditions?*

Conditions We conducted a between-subjects study where participants were randomly assigned to one of four conditions:

- NoGR: Participants made their decisions with no goal recognition output received.
- GR: Participants made their decisions with goal recognition output received.
- XGR_WoE: Participants made their decisions with goal recognition output received along with explanations of our XGR model based on WoE only.
- XGR_WoE_CF: Participants made their decisions with goal recognition output received along with explanations of our XGR model based on both WoE and counterfactual explanations.

We have the NoGR condition to understand the impact of AI assistance on decision-making processes. In addition, we decompose our model into XGR_WoE, where the explanation is generated based on observational markers only, and XGR_WoE_CF, where explanations are generated based on both observational markers and counterfactuals, to facilitate comparison.

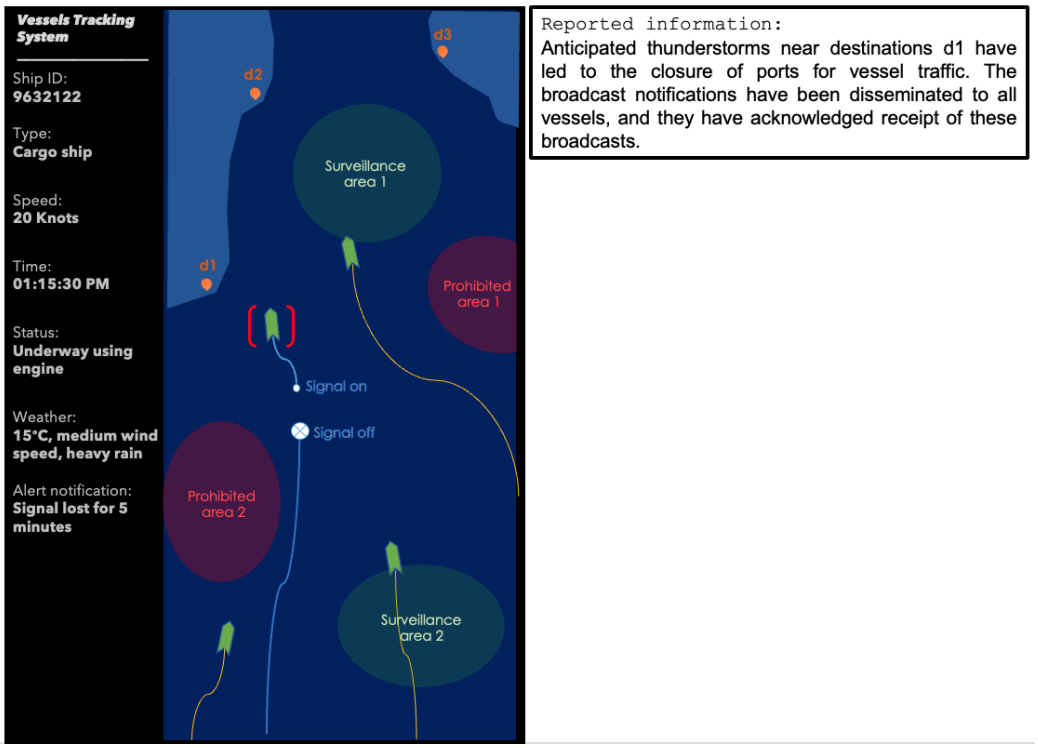


Figure 13: Maritime Domain, Scenario 2 (NoGR Condition).

Task Setup This time we opted to test our model on a more complex domain, inspired by real-world applications (Cordner et al., 2017; Rosello, 2020). We used a maritime domain as our test bed and designed eight scenarios to assess whether a vessel was involved in illegal maritime activities. In the eighth scenario, we introduced two detected vessels (8a, 8b), adding a level of complexity in dealing with multiple vessel interactions within the same environment. The scenarios include invading prohibited areas, deliberately avoiding surveillance areas, or concealing illegal operations by turning off signals.

An example scenario is presented in Figure 13, which shows the detected vessel (red brackets) heading toward one of three destinations (port d1, d2, or d3). The GR task consists of 1) an initial state where the vessel is located at the start of its path on the map (depicted as a blue line); 2) a set of several distinct goal hypotheses, which include whether or not the vessel invades prohibited areas, whether or not the vessel is attempting to avoid surveillance areas, is the vessel attempting to conceal, and the destination it is heading to; and 3) the observed behavior sequence, which includes the vessel's path (represented as a blue line). The vessels navigate within a 2D environment, taking actions such as moving in four directions (up, down, left, right) and toggling their monitoring systems. Key model variables include the vessel's location, the locations of prohibited and surveillance areas, the destinations, and whether the vessel's signal has been lost or if it avoids certain areas. We used a STRIPS-like discrete planner to generate the ground truth plan hypotheses based on the domain theory and observations. Participants were tasked with assigning the likelihood of a detected vessel's possible destination based on its observed behavior, and the likelihood of the vessel being engaged in illegal activity and the need to be intercepted by the Coast Guard. Each participant was exposed to eight scenarios, which were randomly assigned and ordered to avoid order effects. We chose detecting illegal activity as our task since it involves complex, high-stakes scenarios requiring participants to judge based on limited and ambiguous information.

Procedure Participants were presented with eight scenarios of a vessel heading to one of the three destinations (seaports). The experiment consisted of four phases:

- **Phase 1:** Collection of demographic information and participant training. Participants were trained to understand the task using three training scenarios.

- **Phase 2:** Participants were shown a static image simulating the tracking system display, along with the GR system output, referred to as the decision aid system in the second condition. Additionally, the explanation system output was included to answer "why" and "why not" questions in the third and fourth conditions. Participants were then asked to predict the vessel's goals regarding its destination and any potential illegal activities that might require interception by the Coast Guard. Each participant completed eight scenarios. Our algorithm pre-generated the explanations and presented them in natural language (refer to Appendix C3. for example scenarios of the four conditions).

- **Phase 3:** In the last scenario, participants were given an open-ended question to justify their decision.

- **Phase 4:** Completion of the trust scale by participants for all conditions except the first (NoGR).
- **Phase 5 (XGR conditions only):** Completion of the explanation satisfaction scale for explanation conditions (XGR_WoE, XGR_WoE_CF).
- **Phase 6:** After rating explanations, participants were asked two further questions: To what extent did you use the provided information? Options: not at all, minimally, moderately, substantially, extensively. Please briefly describe how you incorporated the presented information into your decision-making process. Participants were encouraged to provide detailed answers. We asked this question to gain insight into their thought processes and to assess their level of engagement with the task.

Participants We conducted a power analysis to determine the required sample size. Using Cohen's F, we obtained an effect size of 0.20. With a power of 0.80 and a significance level of 0.05, we calculated a total sample size of 276 for the four conditions. We recruited 290 participants on Prolific, who were randomly and evenly allocated to each condition. To ensure data quality, participants had to reside in the US, UK, or Australia, be native English speakers, and have a minimum approval rate of 99% with at least 1,000 previous submissions. After excluding inattentive participants, we obtained 280 valid responses (NoGR: 70, GR: 69, XGR_WoE: 72, XGR_WoE_CF: 69). The demographic breakdown included 123 males, 154 females, and 3 self-specified, aged between 18 and 75, with a mean age of 37. Participants were compensated $8.00 USD, with an additional performance-based bonus of up to $3.00 USD.

Metrics To assess the impact of our model on decision accuracy, we used the Brier score function. The Brier score, which ranges from 0 (indicating best performance) to 1 (indicating worst performance), measures the accuracy of predictive probabilities for binary and multiclass outcomes. It is calculated as the mean squared distance between the predicted class probabilities and the actual outcomes (the ground truth) (Brier, 1950):

$$\text{Brier Score} = \sum_{i=1}^{c} (p_i - y_i)^2 \tag{7}$$

Where c is the number of classes, p is the predicted probability, and y is the ground-truth label represented as a vector $y = (y_1, \ldots, y_c)$, where $y_i = 1$ for the true class and 0 otherwise. A lower Brier score indicates better task performance, rewarding high-confidence correct predictions while penalizing high-confidence incorrect ones.

To test the effect of our model on decision-making efficiency, we measured the time participants took to complete each scenario. Since response times can be skewed by outliers, we applied a logarithmic transformation to normalize the data and ensure robust statistical analysis (Vankov, 2023).

To evaluate appropriate reliance, we adopted outcome-based measures of reliance (Ma et al., 2024). This allowed us to distinguish between two key types of misuse:

$$\text{Overreliance} = \frac{\text{Incorrect human decisions with incorrect GR predictions}}{\text{Total number of incorrect GR predictions}} \tag{8}$$

$$\text{Underreliance} = \frac{\text{Incorrect human decisions with correct GR predictions}}{\text{Total number of correct GR predictions}} \tag{9}$$

These metrics provide insights into how well participants calibrated their decisions with the system's predictions.

To test whether our model increases user trust in the GR agent, we used the *trust scale* from Hoffman et al. (2018). Participants rated their trust using a 5-point Likert scale (0 = Strongly Disagree to 100 = Strongly Agree) across four metrics. This hypothesis also served to extend and further validate our prior findings on trust (see Section 6.4.2) in a new task setting. In addition, to assess explanation quality, we employed the *Explanation Satisfaction Scale* from the same source, also using a 5-point Likert scale from 0 to 100 across four evaluative metrics.

Analysis Method Since the collected data did not meet the assumptions required for parametric testing, we used the non-parametric method's Kruskal-Wallis test. Pairwise differences were investigated using the Dunn test with Holm correction. Further, to assess the correlat

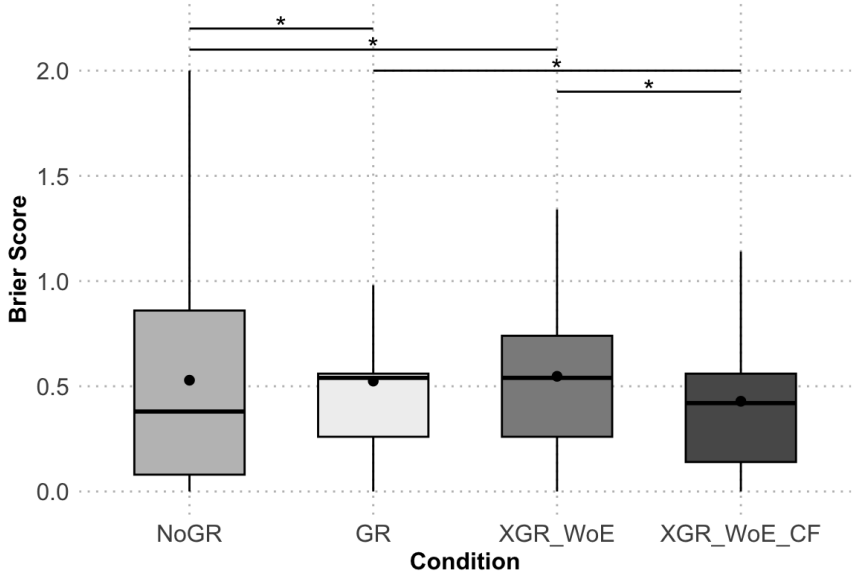


Figure 14: Brier score for the task of vessel destination detection across the four conditions, with means represented as dots (lower is better).

6.6 Results

In this section, we present the results of our experiments, addressing the four hypotheses and the research questions outlined in this study.

6.6.1 Decision Accuracy

We first discuss the results of the first hypothesis, where we investigate whether our model leads to a more accurate decision.

The vessel destination The result of the statistical test for the vessel destination predictions indicates a significant difference between conditions ($\chi^2 = 41.1$, $p < 0.001$). We then performed a pairwise comparison test and the results further indicate significant differences between conditions (Figure 14). There is a statistically significant difference between the XGR_WoE_CF cohort and the GR (with no explanations) cohort with $p < 0.001$. There is a marginally significant difference between the XGR_WoE_CF cohort and the NoGR cohort (no GR output and no explanations) with $p = 0.1$. Thus, we accept our initial hypothesis that our model XGR_WoE_CF leads to more accurate decisions.

Importantly, we also examined the effect of providing only WoE-based explanations (XGR_WoE) without counterfactuals. The results show that XGR_WoE did not significantly improve decision accuracy compared to GR or NoGR. In fact, participants in the XGR_WoE condition performed worse than those in both the GR and NoGR conditions. The integration of counterfactuals (CF) with WoE appears to be essential for enhancing user understanding and decision accuracy. This trend was consistent across all the experiments, and we highlight it as a key finding of this work.

Also, we observed a statistically significant difference between participant performance in the GR and NoGR cohorts with $p = 0.002$ indicating that participants performed worse with GR compared to NoGR. It has been argued that in some cases, people might perform worse with computer support even if the tool offers correct decisions. This can be due to various factors such as their skills and the computer's design (Alberdi et al., 2009). These results are reflected by measuring the appropriate reliance on GR model: while GR appears to help with underreliance, this benefit is countered by an increase in overreliance (Figure 17). This finding further strengthens previous findings (Vered et al., 2023) and suggests that while GR may reduce underreliance, it simultaneously leads to excessive trust in the GR outputs

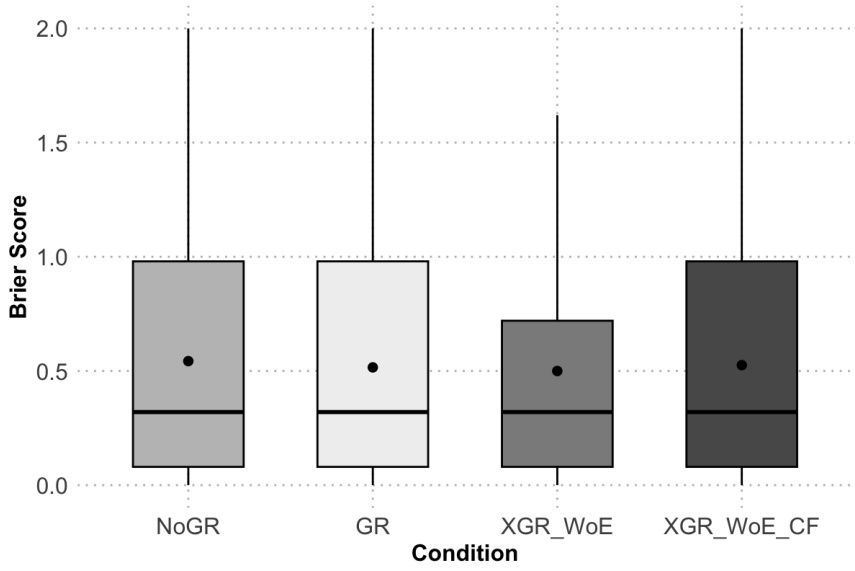


Figure 15: Brier score for the task of dispatching the coast guard, across the four conditions with means represented as dots (lower is better).

Dispatching the Coast Guard When determining whether there is a need to dispatch the Coast Guard to intercept the vessel, the statistical analysis resulted in no statistically significant difference in the Brier scores among the different conditions ($\chi^2 = 0.18$, $p = 0.98$). Figure 15 illustrates these findings using a box plot. Consequently, we reject our hypothesis for this task that our model leads to more accurate decisions.

We hypothesize that the lack of significant difference may be attributed to the inherent complexity of the task, which involves processing multiple streams of information. Participants need to consider vessel behaviors such as invading prohibited areas, avoiding surveillance zones, and concealing activities. The cognitive load required to manage and synthesize all this information could lead to cognitive overload, potentially causing participants to rely less on analytical processing. This aligns with the notion that individuals are cost-sensitive decision-makers who weigh the effort required to use a decision aid system against its potential bene

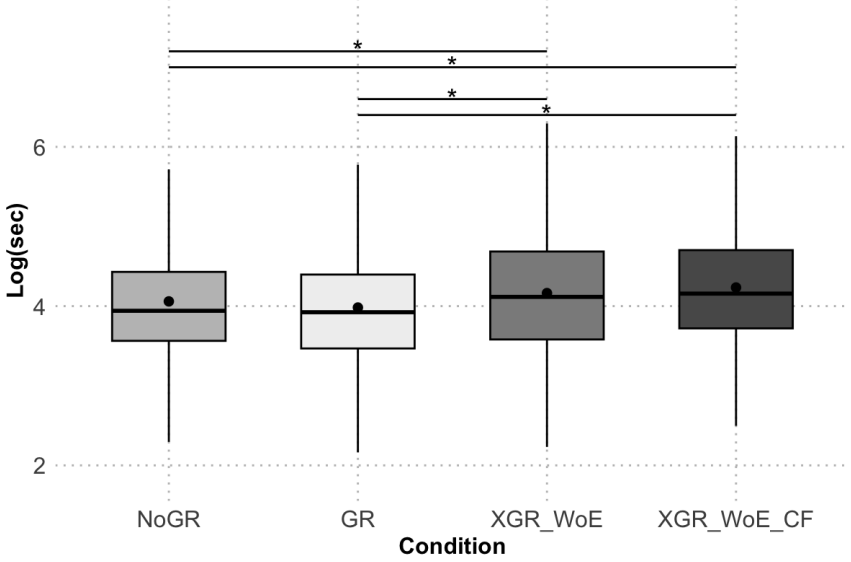


Figure 16: Completion time across the four conditions with means represented as dots (lower is better).

6.6.2 Decision Efficiency

For the second hypothesis, we evaluate how quickly participants made their decisions, hypothesizing that our model will contribute to greater overall efficiency in task completion. Efficiency was measured by the overall time spent on the tasks, recorded in seconds. The statistical analysis revealed significant differences in completion time across the four conditions ($\chi^2 = 39.28$, $p < 0.001$). Post hoc pairwise comparisons revealed that participants in the explanation cohorts, XGR_WoE and XGR_WoE_CF, had significantly higher completion time compared to participants in the NoGR and GR cohorts ($p < 0.001$). As shown in Figure 16, there are significant differences in completion time across the four conditions. However, no statistically significant difference was found between NoGR and GR ($p = 0.13$),

nor between XGR_WoE and XGR_WoE_CF ($p = 0.15$). Therefore, we reject the hypothesis that our models improve overall task efficiency.

Correlation of Decision Accuracy and Efficiency We further examined the relationship between Brier scores and task completion time to understand the influence of time spent on decision accuracy. Our findings suggest that longer task completion times were associated with lower Brier scores, indicating more accurate decisions. This is evidenced by a weak negative correlation between these variables. Notably, the correlation is statistically significant ($p < 0.05$) for certain conditions, as shown in Table 9 for the vessel destination task and Table 10 for the Coast Guard task. These results highlight the importance of allocating sufficient time for information processing to improve decision-making accuracy.

Condition	Correlation	P-Value	Confidence Interval (95%)
NoGR	-0.0470	0.267	[-0.129, 0.0360]
GR	**-0.100**	**0.0183**	**[-0.182, -0.0171]**
XGR_WoE	**-0.179**	**0.0000158**	**[-0.257, -0.0986]**
XGR_WoE_CF	**-0.157**	**0.000222**	**[-0.237, -0.0741]**

Table 9: Pearson correlation coefficient between Brier score and completion time for vessel destination task. Significant values are indicated in bold.

Condition	Correlation	P-Value	Confidence Interval (95%)
NoGR	-0.0600	0.156	[-0.142, 0.0230]
GR	**-0.129**	**0.00234**	**[-0.210, -0.0463]**
XGR_WoE	-0.0609	0.144	[-0.142, 0.0209]
XGR_WoE_CF	**-0.124**	**0.00363**	**[-0.205, -0.0406]**

Table 10: Pearson correlation coefficient between Brier score and completion time for Coast Guard task. Significant values are indicated in bold.

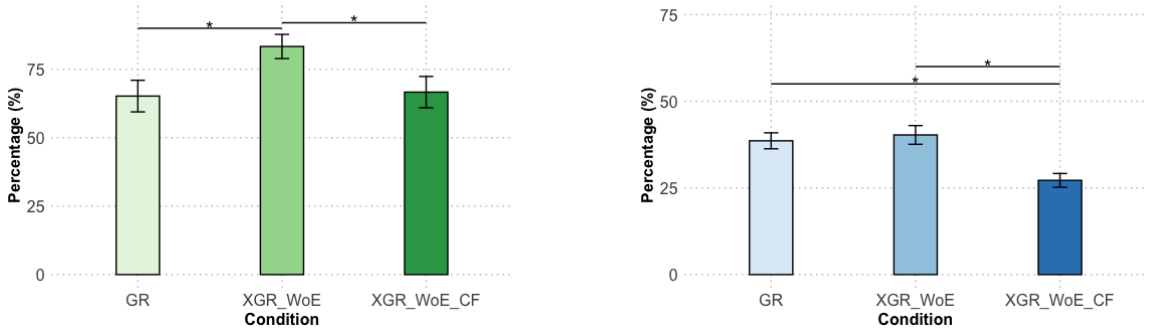
6.6.3 Appropriate Reliance on GR Model

For the third hypothesis, we examine whether our model influences the appropriateness of human reliance, specifically by reducing overreliance and underreliance on the GR model. We measure overreliance as the fraction of incorrect human decisions that align with incorrect GR predictions, calculated over the total number of incorrect GR predictions. Conversely, underreliance is measured as the fraction of incorrect human decisions that occur despite correct GR predictions, calculated over the total number of correct GR predictions.

Vessel destination prediction task. For the task of predicting the vessel's destination, the statistical analysis revealed significant differences between the conditions ($\chi^2 = 7.08$, $p = 0.03$) in terms of **overreliance**. Figure 17(a) shows the results of the pairwise comparisons, indicating significant differences between GR paired with XGR_WoE ($p = 0.02$) and XGR_WoE paired with XGR_WoE_CF ($p = 0.03$). Explanations increased the chance that humans would rely on system predictions (Bansal et al., 2021), but the counterfactual

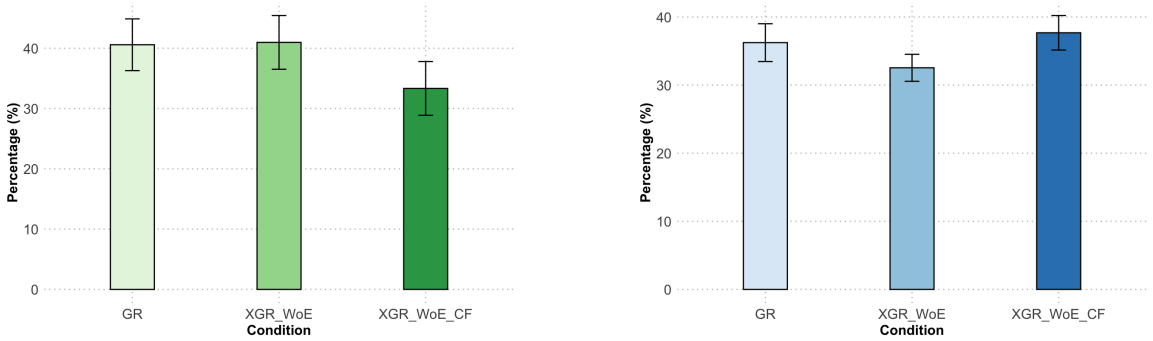
explanation in our model (XGR_WoE_CF) helped reduce that overreliance. Therefore, we accept our hypothesis only for XGR_WoE_CF.

The statistical analysis further revealed significant differences between the conditions ($\chi^2 = 16.1$, $p < 0.001$) in terms of measuring **underreliance**. Figure 17(b) shows the results of the pairwise comparisons, indicating significant differences between GR paired with XGR_WoE_CF ($p < 0.001$) and XGR_WoE paired with XGR_WoE_CF ($p < 0.001$). XGR_WoE_CF significantly reduces underreliance compared to GR and XGR_WoE. We again accept our hypothesis only for XGR_WoE_CF.



(a) Overreliance Results (b) Underreliance Results

Figure 17: Appropriate human reliance for the task of vessel destination prediction (lower is better). Error bars indicate standard errors of the mean.

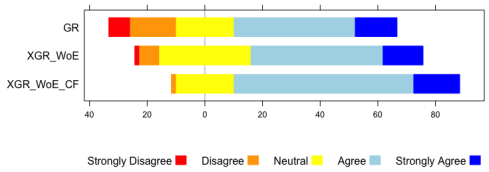


(a) Overreliance Results (b) Underreliance Results

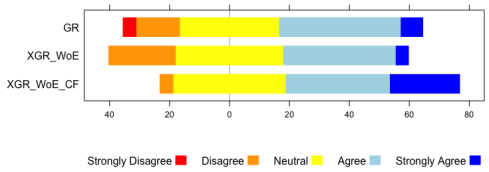
Figure 18: Approriate human reliance for the task of dispatching the coast guard (lower is better). Error bars indicate standard errors of the mean.

Dispatching the Coast Guard We also measured over and under reliance of users on the GR model for the task of deciding whether the Coast Guard should be dispatched to intercept the vessel. Despite a difference in the mean values between the three conditions, no statistically significant differences were found following the statistical analysis ($\chi^2 = 2.12$, $p = 0.35$) to measure the **overreliance** (see Figure 18 (a)). Thus, we reject our hypothesis.

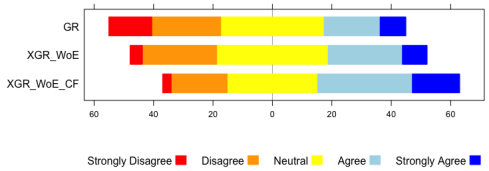
The same result was found when measuring the effect of reducing **underreliance** (see Figure 18 (b)). Despite different mean values between the three conditions, no statistically significant differences were found following the statistical analysis ($\chi^2 = 1.38$, $p = 0.50$). We therefore reject our hypothesis of promoting appropriate reliance in this task. While our model XGR_WoE_CF appears to help with overreliance, this effect is countered by the result of underreliance, which reflects the accuracy results we observed (Figure 15). When people find the task and explanation cognitively demanding, they tend to rely on the system's prediction without carefully verifying it (Vasconcelos et al., 2023).



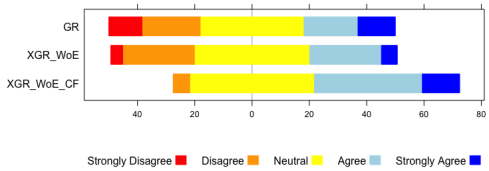
(a) Confident



(b) Predictable



(c) Reliable



(d) Safe

Figure 19: Likert scale of perceived trust metrics across the three conditions. The X-axis represents each Likert category's total counts of responses, adjusted to have 0 as the midpoint.

Table 11: Pairwise comparisons with Post-hoc test for trust metrics

Metric	Comparison	Z	P.adj
Confident	GR vs XGR_WoE	0.11	1.00
	GR vs XGR_WoE_CF	-3.06	<**0.01**
	XGR_WoE vs XGR_WoE_CF	-3.20	<**0.01**
Predictable	GR vs XGR_WoE	0.94	1.00
	GR vs XGR_WoE_CF	-2.47	**0.03**
	XGR_WoE vs XGR_WoE_CF	-3.44	<**0.01**
Reliable	GR vs XGR_WoE	-1.13	0.25
	GR vs XGR_WoE_CF	-3.22	<**0.01**
	XGR_WoE vs XGR_WoE_CF	-2.12	0.06
Safe	GR vs XGR_WoE	0.60	0.55
	GR vs XGR_WoE_CF	-2.89	<**0.01**
	XGR_WoE vs XGR_WoE_CF	-3.52	<**0.01**

6.6.4 User Trust

We now report the results of our evaluation to determine whether our explanation model promotes trust in the GR agent. Figure 19 illustrates the distribution of Likert scale data across conditions. The statistical test was performed, yielding a significant result ($p < 0.001$) for all metrics, indicating a significant difference among the three conditions for the Confident, Predictable, Reliable, and Safe metrics. Subsequently, we conducted pairwise comparison tests, and the results are summarized in Table 11. These findings highlight that the XGR_WoE_CF condition shows significant results across all metrics. Therefore, we accept the hypothesis that our model, XGR_WoE_CF, effectively promotes trust in the GR agent.

Table 12: Mean (SD) of explanation quality across conditions

Condition	Complete	Satisfying	Sufficient Detail	Understand
XGR_WoE	64.2 (21.4)	65.8 (21.8)	67.3 (22.0)	69.2 (19.8)
XGR_WoE_CF	67.5 (19.7)	69.6 (17.4)	71.0 (18.4)	73.6 (15.2)

6.6.5 Explanation Satisfaction

We now report the results of the self-reported metrics of explanation satisfaction. We performed statistical tests to examine whether there are any significant differences between the explanation models XGR_WoE and XGR_WoE_CF for the explanation quality metrics: understanding, satisfying, sufficient detail, and completeness. The obtained p-values $(0.20, 0.31, 0.34, 0.24)$ for these four metrics indicate no significant differences between the two models. Although the differences are not statistically significant, Table 12 shows that the XGR_WoE_CF model is generally better perceived across these metrics, with more consistent responses from participants. This mirrors the results in decision accuracy, as participants feel they have a better understanding of the agent.

6.6.6 Reasoning Variations Across Different Conditions (Research Question 1)

Participants answered an open-ended question about their decision justification in the last scenario. After excluding answers with fewer than three words or containing gibberish, a total of 244 textual data points were collected. We focused on understanding the reasoning process and explanation concepts used across the four conditions. We systematically coded the responses following a hybrid approach of deductive and inductive thematic analysis approach (Braun & Clarke, 2006), labeling responses based on our pre-established main concepts (*observational marker and counterfactual observational marker*). We then analyzed participants' responses and identified a few differences in their justifications across the four conditions.

Explaining Using Observational Markers. We found that across conditions, participants in their justifications mainly relied on the most important evidence for their goal hypothesis (i.e., *observational markers*) (see Figure 20). From the given information, they

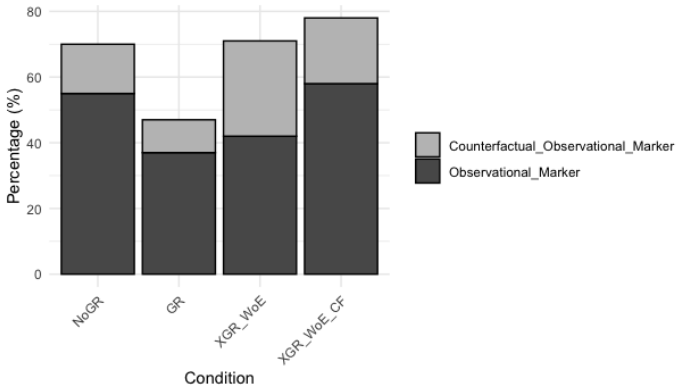


Figure 20: Frequency of using Observational Markers across conditions.

focused only on what they believed were the key features that increased the likelihood of their hypothesis. In the explanation conditions (XGR_WoE and XGR_WoE_CF), they were also able to reason about the *counterfactual observational marker* more frequently, which is evidence against their hypothesized counterfactual goal. This suggests that the explanations, even without explicit counterfactuals (CF), prompted people to think in a counterfactual way, compared to just giving the goal. An example from the data corpus (NoGR condition): "*The speed change of vessel B along with signal loss needed investigating.*".

GR Output Initiates Reasoning. Due to the presence of GR output in the GR, XGR_WoE, and XGR_WoE_CF conditions, we observed that participants often began their justifications by referencing the GR output, explicitly stating their agreement or disagreement. This suggests that the GR output may serve as a reference point to initiate their reasoning. For example, in the GR condition, one participant stated, "*Because the intention to hide was above 50 percent, it would be best if the vessel was intercepted*". Similarly, in the XGR_WoE_CF condition, another participant noted, "*The decision aid (GR system) has made an error in suggesting there is a 100% probability that they have invaded a prohibited area when they were avoiding a prohibited area. This vessel does not require interception*".

Explanations can help ensure that GR output is trustworthy with sufficient certainty. Some participants made efforts to ensure that GR output could be trusted by explicitly discussing their perceptions of the reliability of the GR output to decide how best to use it. They used the provided information to guide their trust in GR. This was particularly evident in conditions with sufficient confidence in XGR_WoE and XGR_WoE_CF. An example from data corpus (XGR_WoE_CF condition): "*Vessel B's signal is lost for 30 minutes near a prohibited area, high probability of intentional concealment, also supported by the decision aid (GR system), and hence Coast Guard should be alerted*".

6.7 Discussion

Our three studies demonstrated the validity of our XGR model. We found that the model enhanced transparency and understanding of GR agents. Additionally, we found that the

counterfactual explanation was a crucial component of our model when addressing "why not" questions, as evidenced in the third study (see results of XGR_WoE_CF, Section 6.5). Counterfactual explanations were preferred by users since they offered coherent and extensive explanations that encompassed multiple instances, aligning with human preferences for comprehensive insights (Miller, 2019b).

However, our findings in the context of decision-making support, the task of dispatching the Coast Guard, were inconclusive. This could arise from a combination of factors, including task difficulty, participant engagement, or the possibility that the explanations provided were not sufficiently useful or aligned with the decision-making requirements of the task. The task complexity may have been beyond the participants' capabilities or understanding, leading to challenges in accurately making decisions. Difficult tasks can result in increased cognitive load and errors, causing people to struggle with accurately evaluating the AI model's output, potentially obscuring the model's true effectiveness (Vasconcelos et al., 2023). This suggests that the cost of verifying the AI model's output in this challenging task may have been so high that it outweighed the benefits, potentially making manual task completion more efficient.

Furthermore, the complexity of the task may have influenced participant engagement. Participants who found the task too challenging might have been less engaged, impacting their performance. Engagement issues could also be linked to the task's perceived relevance and length. Low engagement can result in superficial responses and a lack of attention to detail, as evidenced by some participants completing the tasks very quickly. These observations indicate that both task difficulty and participant engagement played a role in the inconclusive results, suggesting that further investigation is necessary.

7. Conclusion

We developed an explainable model for GR agents that can generate explanations to answer why and why-not questions. The model is grounded in empirical data from two different human-agent studies. We evaluated our model computationally across eight online GR benchmark domains. Additionally, we conducted three human studies to investigate how the XGR model generates human-like explanations, increases user understanding and trust in GR agents, and improves the decision-making process.

While results indicate a significantly better performance of our model compared to the baselines, further research is needed to address the impact of task difficulty and participant engagement on decision-making support. This includes designing appropriately challenging yet understandable tasks and ensuring that participants are adequately motivated and engaged throughout the study. Additionally, we plan to extend our model to handle scenarios with partial observability, where some information may be missing or hidden. Another avenue for future work is to evaluate our model by incorporating additional concepts and examining how they affect user performance and satisfaction.

Acknowledgments

This material is based on research partially sponsored by the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) program under award number FA8750-23-2-1016.

Appendix A. Human-Agent Study 1: Sokoban Game

The Sokoban game is a single-player game where the player is responsible for pushing a box to one of the goal locations on the map, which are identified by numbers. Across fifteen Sokoban game scenarios (five for each game version), the participant's task is to predict which goal the player is pushing the box toward, based on the player's observed behavior.

Below is a screenshot of example scenarios across the three game versions and the corresponding participant task. All data will be made available upon request.

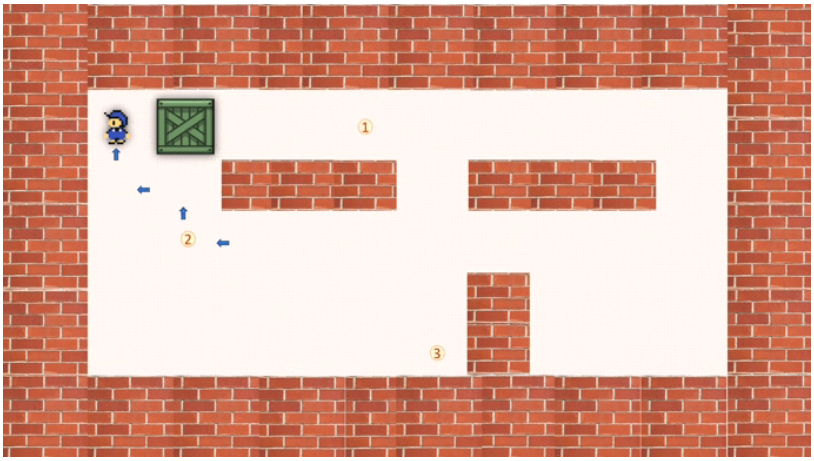


Figure 21: Example scenario from Game Version 1: The player's task is to move one box to the goal location.

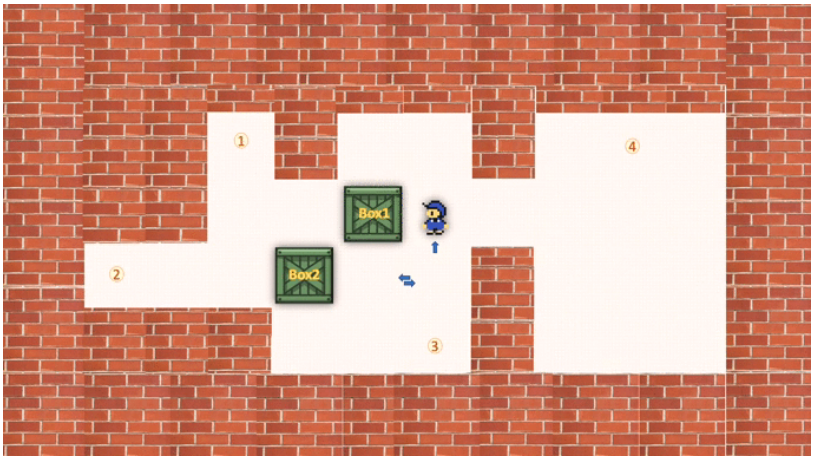


Figure 22: Example scenario from Game Version 2: The player's task is to move two boxes to their designated goal locations, with the ability to push only one box at a time.

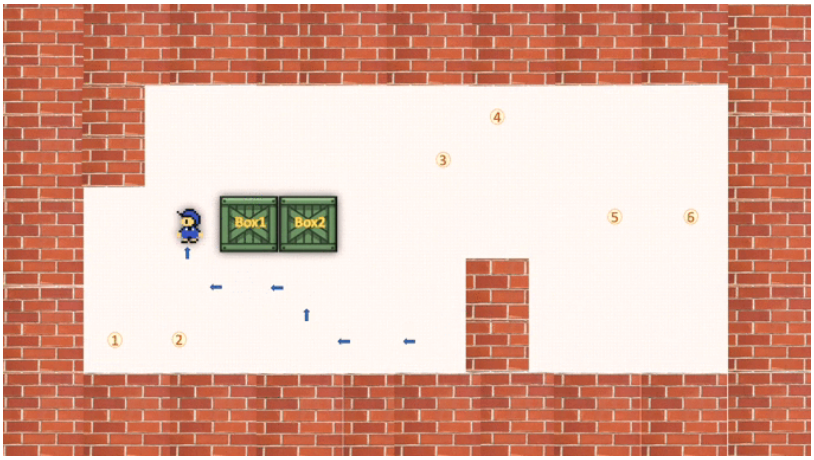


Figure 23: Example scenario from Game Version 3: The player's task is to move two boxes to their designated goal locations, with the ability to push them simultaneously.

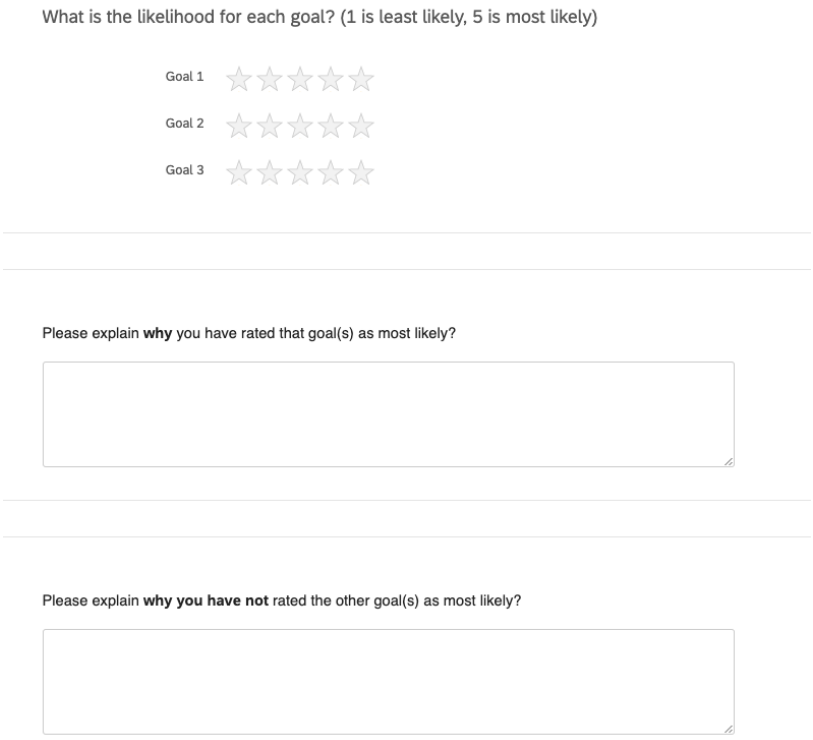


Figure 24: Participant's task from the Dual Condition (Game Version 1): Assign likelihoods to goal locations and answer questions about why and why not.

Appendix B. The Weight of Evidence (WoE): Formula Derivation

The WoE is defined for some evidence e, a hypothesis h, and its logical complement $\overline{h}$ (Good, 1985) as follows:

$$woe(h:e) = \log \frac{Odds(h \mid e)}{Odds(h)} \tag{10}$$

where the colon is read as "provided by", and $Odds(.)$ denotes the hypothesis odds:

$$Odds(h \mid e) = \frac{P(h \mid e)}{P(\overline{h} \mid e)} \quad \text{(Posterior odds)} \tag{11}$$

$$Odds(h) = \frac{P(h)}{P(\overline{h})} \quad \text{(Prior odds)} \tag{12}$$

This is the ratio of the posterior to the prior odds. The odds corresponding to a probability p are defined as $p/(1-p)$ — the probability of an event occurring divided by the probability of it not occurring.

Using Bayes' rule, $woe(h:e)$ can also be defined as:

$$woe(h:e) = \log \frac{P(e \mid h)}{P(e \mid \overline{h})} \tag{13}$$

WoE can also contrast h to an arbitrary alternative hypothesis h' instead of its complement $\overline{h}$. Thus, we can talk generally about the strength of evidence in favor of h and against h' provided by e:

$$woe(h/h':e) = woe(h:e \mid h \vee h') \tag{14}$$

The WoE generalizes to include cases when it can be conditioned on additional information c:

$$woe(h:e \mid c) = \log \frac{P(e \mid h,c)}{P(e \mid h',c)} \tag{15}$$

From various properties of WoE, the following two properties are essential to our model:

$$woe(h/h':e) = \log \frac{P(e \mid h)}{P(e \mid h')} \tag{16}$$

$$\log \frac{P(h)}{P(h')} + \log \frac{P(e \mid h)}{P(e \mid h')} = \log \frac{P(h \mid e)}{P(h' \mid e)} \tag{17}$$

By substituting using Equation 16, we get:

$$\log \frac{P(h)}{P(h')} + woe(h/h':e) = \log \frac{P(h \mid e)}{P(h' \mid e)} \tag{18}$$

$$woe(h/h':e) = \log \frac{P(h \mid e)}{P(h' \mid e)} - \log \frac{P(h)}{P(h')} \tag{19}$$

Using the log quotient property, we simplify the equation as follows:

$$woe(h/h' : e) = \log \frac{\frac{P(h|e)}{P(h'|e)}}{\frac{P(h)}{P(h')}} \tag{20}$$

If we have uniform prior probabilities, we can simplify further and compute WoE for a pair of hypotheses (conditioned on c) as follows:

$$woe(h/h' : e \mid c) = \log \frac{P(h \mid e, c)}{P(h' \mid e, c)} \tag{21}$$

Appendix C. Empirical Evaluation

In this section, we provide a brief description of the tasks and screenshots of selected scenarios from the three conducted studies. All data will be made available upon request.

C1. Study 1 - Generating Human-Like Explanations

Fifteen Sokoban game scenarios (five for each game version) are presented for the goal recognition (GR) agent to determine the most likely goal based on the observed sequence of actions. Given the recognized goal set G and the counterfactual goal set G' at a specific time step, the participant's task involves answering the following questions for each scenario by annotating the map:

- For the 'why' question: *Annotate* the most important action that justifies why the goal is considered the most likely, and the second most important action, if any.
- For the 'why not' question:
    1. *Annotate* the most important action that justifies why the goal is considered less likely, and the second most important action, if any.
    2. Annotate the counterfactual action that should have occurred instead.

Below is a screenshot of an example scenario from Game Version 1, along with the corresponding participant task and an example answer.

C2. Study 2 - Perceived Explainability

In this study, we provided six short gameplay videos of a Sokoban game (two scenarios for each game version). For the GR condition, we presented the output of the Goal Recognition system for each scenario. For the XGR condition, we provided both the GR output and an explanation of this output that addresses two questions: 'Why?' and 'Why not?'

Below is a screenshot showing an example scenario from both conditions, along with the corresponding participant task.

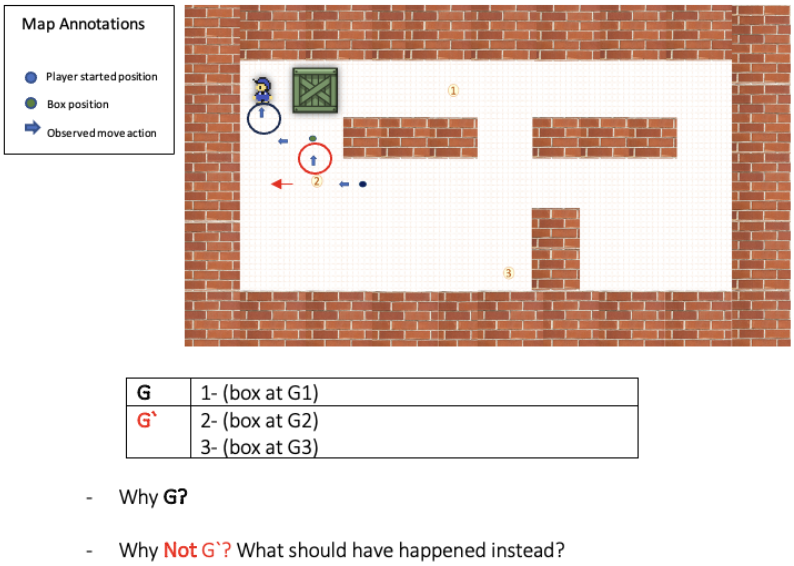


Figure 25: Example scenario from Game Version 1. The black annotation addresses the question 'Why G?' while the red annotation addresses the question 'Why not G'?'

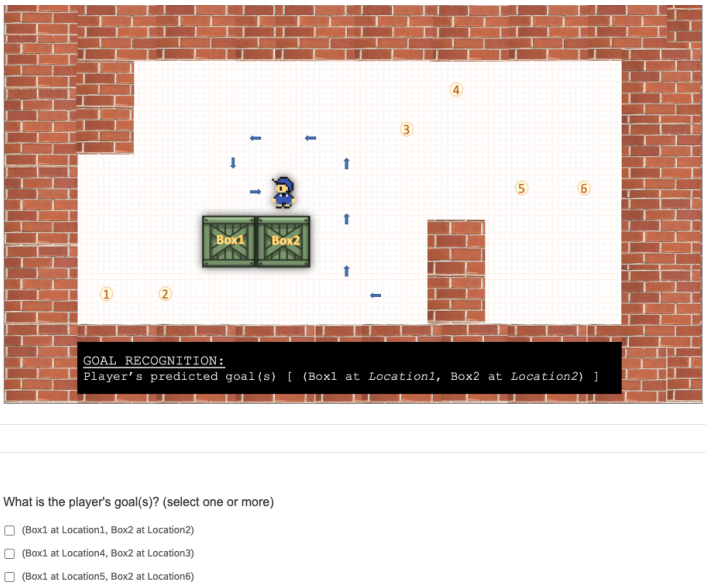


Figure 26: Example scenario from Game Version 3 (GR Condition), where the participant's task is to predict the player's goal locations.

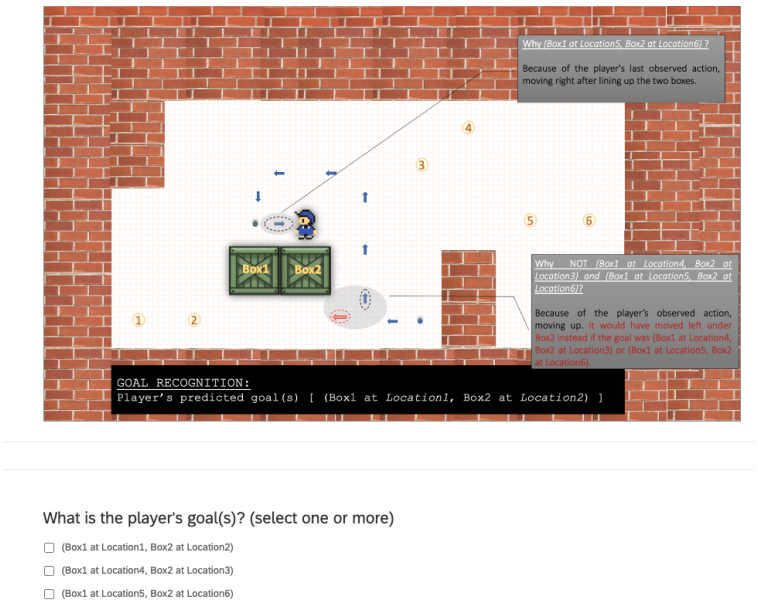


Figure 27: Example scenario from Game Version 3 (XGR Condition), where the participant's task is to predict the player's goal locations.

C3. Study 3 - Effectiveness in Supporting Decision-Making

In maritime surveillance, control centers monitor vessels to detect illegal fishing activities. The main task is to decide on the interception of vessels engaging in such activities, which may involve deploying the Coast Guard. The ocean is divided into prohibited areas, where fishing is forbidden due to conservation concerns, and surveillance areas, where vessels must report their position and activities to authorities. Illegal vessels may either invade prohibited areas or avoid surveillance areas to avoid detection. Additional challenges include potential signal interruptions, which may be due to weak coverage or bad weather, rather than illegal activities.

The participant's task is to incorporate the evidence provided in the scenario to make a final decision regarding the vessel's destination and the necessity of dispatching the Coast Guard. The likelihood of a vessel engaging in illegal activities increases significantly when multiple pieces of evidence are highly predicted, including the intention to invade prohibited areas; the intention to avoid surveillance areas; and the intention to conceal activities.

Below is a screenshot showing an example scenario from all conditions, along with the corresponding participant tasks.

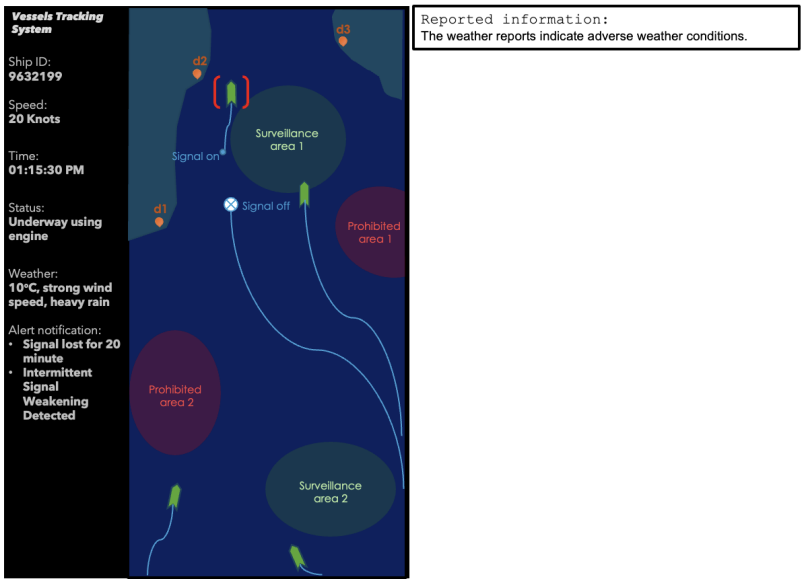


Figure 28: Example scenario (NoGR Condition).

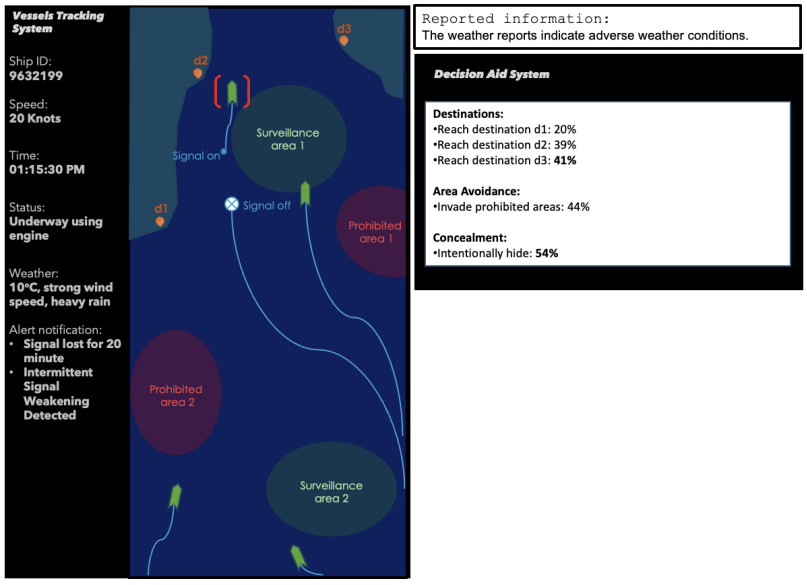


Figure 29: Example scenario (GR Condition).

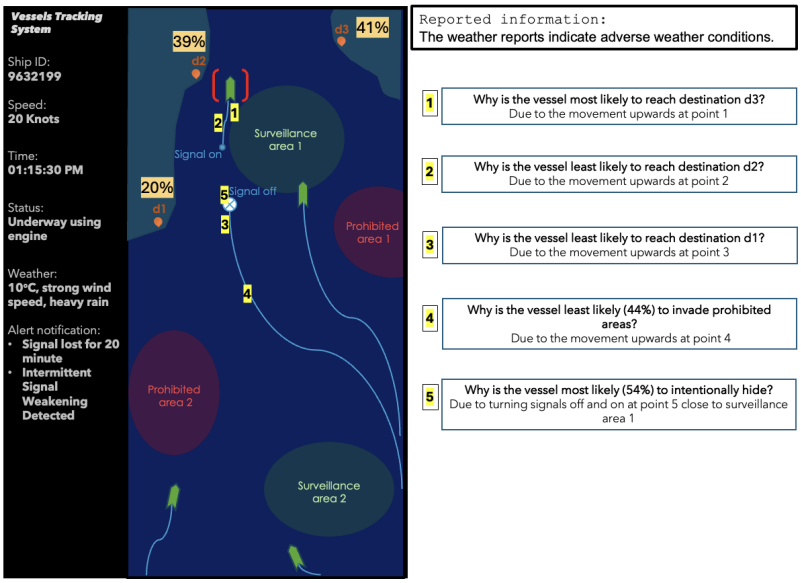


Figure 30: Example scenario (XGR_WoE Condition).

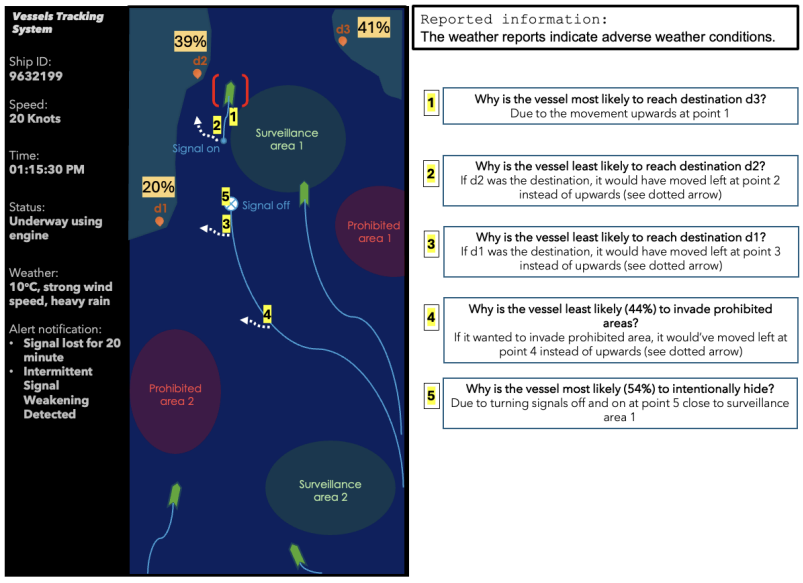


Figure 31: Example scenario (XGR_WoE_CF Condition).

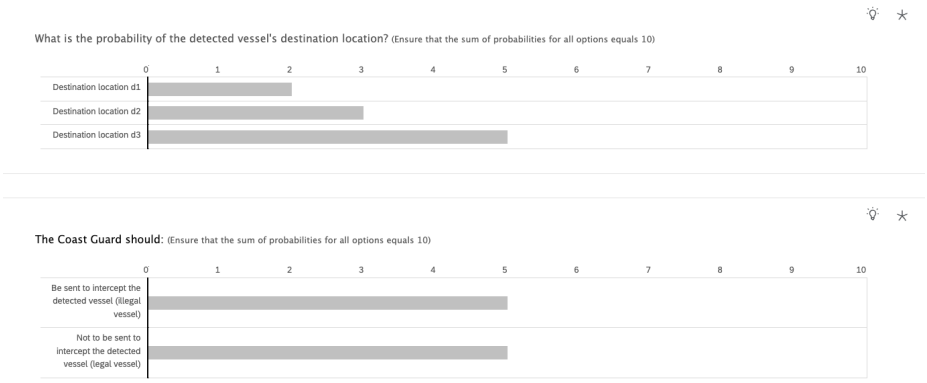


Figure 32: Participant task, rate the likelihood of the vessel's destination and the necessity of intercepting the vessel on a scale from 0 to 10 (where 10 is most likely).

References

Alberdi, E., Strigini, L., Povyakalo, A. A., & Ayton, P. (2009). Why are people's decisions sometimes worse with computer support? *Computer Safety, Reliability, and Security: 28th International Conference, SAFECOMP 2009, Hamburg, Germany, September 15-18, 2009. Proceedings 28*, 18–31.

Alshehri, A., Miller, T., & Vered, M. (2023). Explainable goal recognition: A framework based on weight of evidence. *Proceedings of the International Conference on Automated Planning and Scheduling*, *33*, 7–16.

Amado, L., Mirsky, R., & Meneguzzi, F. (2022). Goal recognition as reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(9), 9644–9651.

Araujo, J., & Born, D. G. (1985). Calculating percentage agreement correctly but writing its formula incorrectly. *The Behavior Analyst*, *8*(2), 207.

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. *Proceedings of the thirtieth annual conference of the cognitive science society*, 1447–1452.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baker, C. L. (2012). *Bayesian theory of mind: Modeling human reasoning about beliefs, desires, goals, and social relations* (Doctoral dissertation). Massachusetts Institute of Technology.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–16.

Bertossi, L. (2020). An asp-based approach to counterfactual explanations for classification. *International Joint Conference on Rules and Reasoning*, 70–81.

Blokpoel, M., Kwisthout, J., van der Weide, T. P., Wareham, T., & van Rooij, I. (2013). A computational-level explanation of the speed of goal inference. *Journal of Mathematical Psychology*, *57*(3-4), 117–133.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101.

Brewitt, C., Gyevnar, B., Garcin, S., & Albrecht, S. V. (2021). Grit: Fast, interpretable, and verifiable goal recognition with learned decision trees for autonomous driving. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1023–1030.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Chiari, M., Gerevini, A. E., Percassi, F., Putelli, L., Serina, I., & Olivato, M. (2023). Goal recognition as a deep learning task: The grnet approach. *Proceedings of the International Conference on Automated Planning and Scheduling*, *33*, 560–568.

Cohen, P. R., & Galescu, L. (2023). A planning-based explainable collaborative dialogue system. *CoRR*, *abs/2302.09646*. https://doi.org/10.48550/ARXIV.2302.09646

Cordner, L., Cordner, L., & Roughley. (2017). *Maritime security risks, vulnerabilities and cooperation*. Springer.

De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposium Series*.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*(1), 32–64.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and social psychology review*, *12*(2), 168–192.

Farrell, R., & Ware, S. (2020). Narrative planning for belief and intention recognition. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *16*(1), 52–58.

Felli, P., Miller, T., Muise, C., Pearce, A. R., & Sonenberg, L. (2015). Computing social behaviours using agent models. *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Fitzpatrick, G., Lipovetzky, N., Papasimeon, M., Ramirez, M., & Vered, M. (2021). Behaviour recognition with kinodynamic planning over continuous domains. *Frontiers in Artificial Intelligence*, *4*, 717003.

Geffner, H., & Bonet, B. (2022). *A concise introduction to models and methods for automated planning*. Springer Nature.

Good, I. J. (1985). Weight of evidence: A brief survey. *Bayesian statistics*, *2*, 249–270.

Hanna, J. P., Rahman, A., Fosong, E., Eiras, F., Dobre, M., Redford, J., Ramamoorthy, S., & Albrecht, S. V. (2021). Interpretable goal recognition in the presence of occluded factors for autonomous vehicles. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7044–7051.

Hegde, R. M., & Kenchannavar, H. H. (2019). A survey on predicting resident intentions using contextual modalities in smart home. *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC)*, *11*(4), 44–59.

Heider, F. (1958). The naive analysis of action. *The Psychology of Interpersonal Relations*. https://api.semanticscholar.org/CorpusID:222503894

Hempel, C. G. (1961). Rational action. *Proceedings and Addresses of the American Philosophical Association*, *35*, 5–23.

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68–73.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *ArXiv*, *abs/1812.04608*. https://api.semanticscholar.org/CorpusID:54577009

Horgan, T., & Woodward, J. (2013). Folk psychology is here to stay. In *Folk psychology and the philosophy of mind* (pp. 144–166). Psychology Press.

Hu, Y., Xu, K., Subagdja, B., Tan, A.-H., & Yin, Q. (2021). Interpretable goal recognition for path planning with ART networks. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Inam, R., Raizer, K., Hata, A., Souza, R., Forsman, E., Cao, E., & Wang, S. (2018). Risk assessment for human-robot collaboration in an automated warehouse scenario. *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, *1*, 743–751.

Jamakatel, P., Bercher, P., Schulte, A., & Kiam, J. J. (2023). Towards intelligent companion systems in general aviation using hierarchical plan and goal recognition. *Proceedings of the 11th International Conference on Human-Agent Interaction*, 229–237.

Kaminka, G., Vered, M., & Agmon, N. (2018). Plan recognition in continuous domains. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Keren, S., Gal, A., & Karpas, E. (2014). Goal recognition design. *Proceedings of the International Conference on Automated Planning and Scheduling*, *24*, 154–162.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, *28*(9), 1321–1333.

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the SIGCHI conference on human factors in computing systems*, 2119–2128.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, *61*(4), 303–332.

Luo, Y., & Baillargeon, R. (2010). Toward a mentalistic account of early psychological reasoning. *Current directions in psychological science*, *19*(5), 301–307.

Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024). "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20.

Malle, B. F. (2006a). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

Malle, B. F. (2006b). Folk Theory of Mind: Conceptual Foundations of Human Social Cognition. In *The New Unconscious*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195307696.003.0010

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology*, *33*(2), 101–121.

Masters, P., & Sardina, S. (2021). Expecting the unexpected: Goal recognition for rational and irrational agents. *Artificial Intelligence*, *297*, 103490.

Masters, P., & Vered, M. (2021). What's the context? implicit and explicit assumptions in model-based goal recognition. *IJCAI*, 4516–4523.

McClure, J., & Hilton, D. (1997). For you can't always get what you want: When preconditions are better explanations than goals. *British Journal of Social Psychology*, *36*(2), 223–240.

Melis, D. A., Kaur, H., Daumé III, H., Wallach, H., & Vaughan, J. W. (2021). From human explanation to model interpretability: A framework based on weight of evidence. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*, 35–47.

Meneguzzi, F. R., & Pereira, R. F. (2021). A survey on goal recognition as planning. *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), 2021, Canadá.*

Miller, T. (2019a). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Miller, T. (2019b). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007

Min, W., Ha, E. Y., Rowe, J., Mott, B., & Lester, J. (2014). Deep learning-based goal recognition in open-ended digital games. *Tenth artificial intelligence and interactive digital entertainment conference.*

Mohseni, S., Block, J. E., & Ragan, E. D. (2020). A human-grounded evaluation benchmark for local explanations of machine learning. https://arxiv.org/abs/1801.05075

Norling, E. J. (2009). *Modelling human behaviour with bdi agents* (Doctoral dissertation).

Ognibene, D., Mirante, L., & Marchegiani, L. (2019). Proactive intention recognition for joint human-robot search and rescue missions through monte-carlo planning in pomdp environments. *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*, 332–343.

Ontanón, S., Synnaeve, G., Uriarte, A., Richoux, F., Churchill, D., & Preuss, M. (2013). A survey of real-time strategy game ai research and competition in starcraft. *IEEE Transactions on Computational Intelligence and AI in games*, *5*(4), 293–311.

Pearl, J., et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19.*

Penney, S., Dodge, J., Anderson, A., Hilderbrand, C., Simpson, L., & Burnett, M. (2021). The shoutcasters, the game enthusiasts, and the ai: Foraging for explanations of real-time strategy players. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(1), 1–46.

Pereira, R., Oren, N., & Meneguzzi, F. (2017). Landmark-based heuristics for goal recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Pereira, R. F., Oren, N., & Meneguzzi, F. (2020). Landmark-based approaches for goal recognition as planning. *Artificial Intelligence*, *279*, 103217.

Pereira, R. F., Vered, M., Meneguzzi, F. R., & Ramırez, M. (2019). Online probabilistic goal recognition over nominal models. *Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, China.*

Pushp, S., Bhardwaj, B., & Hazarika, S. M. (2017). Cognitive decision making for navigation assistance based on intent recognition. *Mining Intelligence and Knowledge*

Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings 5, 81–89.

Ramırez, M., & Geffner, H. (2011). Goal recognition over pomdps: Inferring the intention of a pomdp agent. *IJCAI*, 2009–2014.

Ramırez, M., & Geffner, H. (2010). Probabilistic plan recognition using off-the-shelf classical planners. *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Rizzolatti, G. (2005). The mirror neuron system and its function in humans. *Anatomy and embryology*, *210*(5), 419–421.

Rosello, M. (2020). Illegal, unreported and unregulated (iuu) fishing as a maritime security concern. In L. Otto (Ed.), *Global challenges in maritime security: An introduction* (pp. 33–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-34630-0_3

Sahoh, B., & Choksuriwong, A. (2023). The role of explainable artificial intelligence in high-stakes decision-making systems: A systematic review. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7827–7843.

Santos, L. R., Meneguzzi, F., Pereira, R. F., & Pereira, A. G. (2021). An lp-based approach for goal recognition as planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(13), 11939–11946.

Shvo, M., & McIlraith, S. A. (2020). Active goal recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(06), 9957–9966.

Singh, R., Miller, T., Newn, J., Velloso, E., Vetere, F., & Sonenberg, L. (2020). Combining gaze and ai planning for online human intention recognition. *Artificial Intelligence*, *284*, 103275.

Sohrabi, S., Riabov, A. V., & Udrea, O. (2016). Plan recognition as planning revisited. *IJCAI*, 3258–3264.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323.

Sukthankar, G., Geib, C., Bui, H., Pynadath, D., & Goldman, R. P. (2014). *Plan, activity, and intent recognition: Theory and practice*. Newnes.

Van Rooij, I., Haselager, W., & Bekkering, H. (2008). Goals are not implied by actions, but inferred from actions and contexts. *Behavioral and Brain Sciences*, *31*(1), 38–39.

Van-Horenbeke, F. A., & Peer, A. (2021). Activity, plan, and goal recognition: A review. *Frontiers in Robotics and AI*, *8*, 643010.

Vankov, I. I. (2023). The hazards of dealing with response time outliers. *Frontiers in Psychology*, *14*, 1220281.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance onAI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38.

Vered, M., & Kaminka, G. (2017). Heuristic online goal recognition in continuous domains. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 4447–4454. https://doi.org/10.24963/ijcai.2017/621

Vered, M., Kaminka, G. A., & Biham, S. (2016). Online goal recognition through mirroring: Humans and agents. *The Fourth Annual Conference on Advances in Cognitive Systems*, *4*.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, *322*, 103952.

Vered, M., Pereira, R. F., Magnaguagno, M. C., Kaminka, G. A., & Meneguzzi, F. (2018). Towards online goal recognition combining goal mirroring and landmarks. *AAMAS*, 2112–2114.

Wayllace, C., Ha, S., Han, Y., Hu, J., Monadjemi, S., Yeoh, W., & Ottley, A. (2020). Dragon-v: Detection and recognition of airplane goals with navigational visualization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), 13642–13643.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Yolanda, E., R-Moreno, M. D., Smith, D. E., et al. (2015). A fast goal recognition technique based on interaction estimates. *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, *33*, 19238–19250.