Proceedings of the Thirty-Third International Conference on Automated Planning and Scheduling (ICAPS 2023)

Explainable Goal Recognition: A Framework Based on Weight of Evidence

Abeer Alshehri[1, 3], **Tim Miller**[1], **Mor Vered**[2]

[1] University of Melbourne, Melbourne, Australia
[2] Monash University, Melbourne, Australia
[3] King Khalid University, Abha, Saudi Arabia
aalshehri@student.unimelb.edu.au, tmiller@unimelb.edu.au, mor.vered@monash.edu

Abstract

We introduce and evaluate an eXplainable Goal Recognition (XGR) model that uses the Weight of Evidence (WoE) framework to explain goal recognition problems. Our model provides human-centered explanations that answer 'why?' and 'why not?' questions. We computationally evaluate the performance of our system over eight different domains. Using a human behavioral study to obtain the ground truth from human annotators, we further show that the XGR model can successfully generate human-like explanations. We then report on a study with 60 participants who observe agents playing Sokoban game and then receive explanations of the goal recognition output. We investigate participants' understanding obtained by explanations through task prediction, explanation satisfaction, and trust.

In recent years, a significant amount of research has been conducted on explainable AI (XAI) to increase the transparency of AI decision-making and improve the user's trust (Vered et al. 2020). Although the main focus has been on Explainable Machine Learning, recently, there has been growing interest in Explainable Agency (Langley et al. 2017; de Graaf et al. 2018; Hoffmann and Magazzeni 2019; Chakraborti, Sreedharan, and Kambhampati 2020); agents and robots capable of explaining their decisions to lay users.

For a goal recognition (GR) problem, the task is to infer the most likely goal given an observed agent's behavior. For instance, when autonomous vehicles' anticipated goals are justified to end-users, it would assist to calibrate their trust in such systems (Shahrdar, Menezes, and Nojoumian 2018). We are motivated by the necessity of generating human-like explanations for why a certain goal is most likely. Model-based GR approaches use domain models to generate plans for goals (Masters and Vered 2021), and machine-learning GR approaches rely on the existence of a corpus of prior plans/observations from which to train (Pereira et al. 2019).While there are many ways to achieve this objective, little to no attention has been given to explaining the output of these algorithms, once achieved.

There is a substantial body of literature in cognitive science that explores how humans explain others' behavior (Kashima, McKintyre, and Clifford 1998; Malle 2006; Heider 2013). People's view of the world is normally characterized by their beliefs, goals, and intentions. Reasoning over these mental states with causal relationships lies at the foundation of folk explanations of human behavior (Malle et al. 2000). In light of the existing theory of behavior explanation (Malle 2006), Alshehri et al. (2021) developed a conceptual framework grounded on a human study, with the aim to learn how people explain GR agent behavior; what concepts people use to generate explanations for answering 'why' and 'why not' questions. However, they have not built an explainability method. In this paper, we extend (Alshehri et al. 2021)'s work and propose an *eXplainable Goal Recognition model (XGR)* model that generates explanations consistent with the corresponding human explanation.

We introduce a general XGR model based on the concept of Weight of Evidence (WoE) from information theory (Good 1985; Melis et al. 2021). The model explains the output goal hypothesis of a GR algorithm by obtaining WoE values of observed behavior to determine to what extent an observation is responsible for one goal hypothesis in contrast to another. We define an explanation selection for 'why goal g?' and 'why not goal g'?' questions based on the concept of observational markers, the observation with the highest WoE, and counterfactual observational markers, the observation with the lowest WoE (Alshehri et al. 2021).

We computationally evaluate our approach on eight GR benchmark domains using a state-of-the-art GR model (Vered et al. 2018). Results indicate that our model's computation time is a fraction of the original GR approach. We also conduct a follow-up human study in which participants were presented with the output of the GR model and asked to answer *Why?* and *Why not?* questions. We evaluate our model by comparing human-generated explanations to the output of the XGR model. Results show the efficiency of our model to generate human-like explanations. We conduct another human study using the proposed model for the GR agent that predicts a player's goal in the Sokoban game. Experiments were run for 60 participants, in which we evaluate the participants' performance in task prediction, explanation satisfaction, and trust. Results show that our model has a better performance than the tested baseline. To the best of our knowledge, this is the first study to solve the problem of GR explainability more naturally and elegantly by adopting the concept of WoE.

Related Work and Background

Explainable Agency

A number of studies focus on generating explanations for action/activity recognition models and domains. Some recent examples include (Meng et al. 2019), which uses LSTM based attention mechanism to identify the most relevant frames for video action recognition, and (Akula et al. 2022) which explains decisions made by a deep CNN over image recognition models. These approaches, as well as others, rely on machine learning to generate explanations rather than having an explicit model of the recognizing agent.

Other approaches, such as Albrecht et al. (2021) and Brewitt et al.; Brewitt, Tamborski, and Albrecht (2021; 2022) rely on the innate interpretability of the structure of their specific GR approach. Brewitt et al. (2021) utilize decision trees trained on vehicle trajectory data and Albrecht et al. (2021) rely on inverse planning and Monte Carlo Tree Search. While not dependent on ML, these approaches also do not perform model-based explanations and are only adapted to one specific instance of a GR algorithm.

In the context of sequential decision-making, several explainable agency models have been proposed for Belief-Desire-Intention (BDI) agents (Cranefield et al. 2017; Winikoff, Dignum, and Dignum 2018), reinforcement learning RL agents (Fukuchi et al. 2017; Madumal et al. 2019), and planning agents (Chakraborti et al. 2017; Chakraborti, Sreedharan, and Kambhampati 2018; Cashmore et al. 2019).These frameworks are mostly driven by goal-directed tasks over understanding the autonomous agents' decisions. These approaches, however, do not focus on GR agents.

Previous studies investigating the explainability of goal/intention recognition agent falls into the scope of maximizing the explicability of the agent behavior (Yolanda et al. 2015; Sohrabi, Riabov, and Udrea 2016; Vered, Kaminka, and Biham 2016; Hu et al. 2021; Hanna et al. 2021). This involves making that behavior more explicable to an observer by either aligning its behavior with the observer's expectations or making its inference formation interpretable.

Planning

Planning is a way to find a sequence of actions (i.e, a plan) that achieves a certain goal from an initial state. The concept of planning is key to understanding GR algorithms that utilize planners in the recognition process. Our *eXplainable Goal Recognition model (XGR)* generates explanations for such GR models and also uses off-the-shelf planners to generate a counterfactual plan as part of that explanation. We build upon the following planning problem definition as defined in (Pereira, Oren, and Meneguzzi 2017):

Definition 1. A planning task is represented by a triple $\langle \Xi, \mathcal{I}, g \rangle$, in which $\Xi = \langle \mathcal{F}, \mathcal{A} \rangle$ is a planning domain definition; $\mathcal{F}$ consists of a finite set of facts and $\mathcal{A}$ is a finite set of actions; $\mathcal{I}$ 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 $\mathcal{I}$ by following a sequence of actions. Since actions have an associated cost, we assume that this cost is 1 for all actions. The objective is to find the optimal plan π^* that minimizes the associated cost.

Goal Recognition (GR)

Goal recognition (GR) is the inverse of the planning problem. It is the task of recognizing an agent's unobserved goal through a sequence of observations. There are many different approaches to solving the GR problem. Among the most common approaches are; library based GR algorithms that use dedicated plan recognition libraries that aim to represent all known ways to achieve known goals (Sukthankar et al. 2014); Model-based GR algorithms (Ramírez and Geffner 2010; Sohrabi, Riabov, and Udrea 2016; Vered, Kaminka, and Biham 2016) in which GR agents use their domain knowledge, represented through the use of planners, to generate plans that must be carried out for a goal to be achieved (Masters and Vered 2021); and machine-learning GR approaches that rely on the existence of a large training corpus from which algorithms can learn about the constraints of the domain (Min et al. 2014; Pereira et al. 2019; Meneguzzi and Pereira 2021; Fitzpatrick et al. 2021). Among all of these approaches little, to no, attention has been given to explaining the reasons behind the predicted goals and/or goal probability distribution, which is the aim of this work. To begin we build on the following formal GR definition as defined by (Shvo and McIlraith 2020).

Definition 2. A goal recognition problem is a tuple $\langle \Xi, \mathcal{I}, \mathcal{G}, \mathcal{O} \rangle$, in which $\Xi = \langle \mathcal{F}, \mathcal{A} \rangle$ is a planning domain definition where $\mathcal{F}$ and $\mathcal{A}$ are sets of facts and actions, respectively; $\mathcal{I}$ is the initial state; $\mathcal{G} = \{g_1, g_2, ..., g_m\}$ is the goals set, and $\mathcal{O} = \langle o_1, o_2, ..., o_n \rangle$ is a sequence of observations such that each o_i is a pair (α_i, ϕ_i) composed of an observed action $\alpha_i \in \mathcal{A}$ and a fact set that represent the state $\phi_i \subseteq \mathcal{F}$. A solution to a GR problem is a probability distribution over $\mathcal{G}$ giving the corresponding likelihood of each goal, i.e. the posterior probability $P(g_j \mid \mathcal{O})$ for each $g_j \in \mathcal{G}$. The most likely goal is the one whose generated plan "best satisfies" the observations.

The Mirroring GR Algorithm As part of our model's empirical evaluation, we will be explaining the output of the *Mirroring* GR algorithm (Vered and Kaminka 2017; Kaminka, Vered, and Agmon 2018). The model has been inspired by humans' ability to do online GR which stems from the human brain's mirror neuron system for matching the observation and execution of actions (Rizzolatti 2005). The approach belongs to the *plan recognition as planning* GR approaches (Masters and Vered 2021) and utilizes a planner within the recognition process to compute alternative plans. In particular, the Mirroring algorithm calls a planner first to pre-compute optimal plans from $\mathcal{I}$ to every $g_j \in \mathcal{G}$ as well as to compute *suffix* plans from the last observation $o_i \in \mathcal{O}$ to every $g_j \in \mathcal{G}$. These *suffix* plans are concatenated with a *prefix* plan (the observation sequence $\mathcal{O}$ at time step t) to generate new plan hypotheses. The algorithm then provides a likelihood distribution (posterior probabilities) over $\mathcal{G}$ by evaluating which of the generated plans, that incorporate the observations $\mathcal{O}$, best matches the optimal plan.

Running Example

To better explain the concepts previously introduced we present an MDP model of a navigational GR scenario,

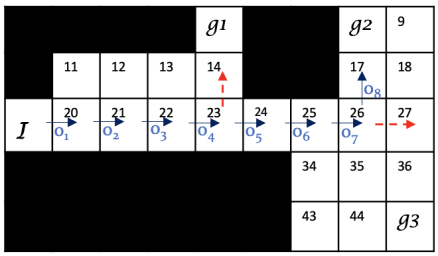

Figure 1: Navigational domain example. $\mathcal{I}$ is the agent's initial state. Given three possible goal locations (g_1, g_2, g_3), the predicted goal is g_2. Blue arrows represent the observation sequence, and red arrows represent counterfactual actions.

shown in Figure 1. In this scenario, an agent must navigate through the unblocked grid to reach one of three possible goal cells, g_1, g_2, or g_3. The state space $\mathcal{S}$ is defined by the cells, 45 states in total, and the initial state of the agent is at cell 19, labeled $\mathcal{I}$. The action set $\mathcal{A}$ comprises moving one cell in each of the four directions (up, down, left, right) with equal cost, and transitions the agent between two connected cells. The GR problem in this example is composed of the following: the initial state, $\mathcal{I}$, set of goal hypotheses, $\mathcal{G} = \{g_1, g_2, g_3\}$, and a sequence of observations $\mathcal{O} = \langle o_1, .., o_8 \rangle$ at time step $t = 8$ whose transition between them is represented as blue arrows. The problem here is deterministic, i.e., at a given state, each action leads to a determined state. As the goal state specification, $\mathcal{G}$ is for the agent to be at one of three possible goal cells. Over this example, the Mirroring GR algorithm would rank goal g_2 as most likely since the observation sequence confirms the optimal plan to achieve this goal. We will refer back to this example throughout the paper to show the performance of our approach on the output of the Mirroring GR algorithm.

Weight of Evidence (WoE)

The principle of rational action (Hempel 2013) states that people explain goal hypotheses by determining to what extent each observed action is responsible for a goal hypothesis in contrast to others. Based on this, Bertossi (2020) defines a causal explanation to be the most responsible set of features for an outcome. Thus, we model our explanation model using the concept of Weight of Evidence.

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

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

[1]Formula derivation is in the supplementary material for when priors are not uniform. To access the full paper see https://arxiv.org/abs/2303.05622

Melis et al. (2021) propose a framework based on WoE for explaining machine learning classification problems and argue that this is a natural model that corresponds to the phenomena of how people explain to each other (Miller 2019). Melis et al. (2021) found WoE naturally captures a contrastive statement, i.e. evidence for or against something. That would help answer questions like why goal g, why not goal g', and what should have happened instead if the goal was g'. We adopt this concept and apply it to GR problems.

eXplainable Goal Recognition (XGR)

In this section, we present a simple and elegant explainability model for GR algorithms called *eXplainable Goal Recognition* (XGR). Extending Melis et al.'s (2021) WoE framework, our model answers 'why' and 'why not' questions, as they are the most demanded explanatory questions (Lim, Dey, and Avrahami 2009).

Our explanation generation approach consists of two parts. The first part ranks each observation in an observed plan by its WoE score. The second part selects highly-ranked observations, obtaining a minimal complete explanation.

Model Overview

Our explainable GR model accepts four inputs, which can be provided by any GR model:

1. An observed sequence $\mathcal{O}$;
2. A set of predicted goals, $G_p \subseteq \mathcal{G}$;
3. A set of counterfactual (not predicted) goals, $G_c \subset \mathcal{G}$, whereby $G_p \cap G_c = \emptyset$; and
4. Posterior probabilities $P(g \mid \mathcal{O})$ for each $g \in \mathcal{G}$.

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.

Explanation Generation

We define an explanation as a pair containing: (1) an explanandum, the event to be explained; and (2) an explanan, a list of causes given as the explanation (Miller 2019). The explanandum is assumed to be of the form *why/why not g?*, where g is a goal. We extend the WoE framework (Melis et al. 2021) to GR problems.

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 with the observation $o_i \in \mathcal{O}$, the additional information c with the observed sequence $\mathcal{O}$ up to the observation o_i, and the posterior probabilities as $P(g \mid \mathcal{O}_i)$ and $P(g' \mid \mathcal{O}_i)$, in which $g \in G_p$ and $g' \in G_c$. We define a complete explanan as follows.

Definition 3. A *complete explanan* for a goal g is a list of pairs $(woe(g/g' : o_i \mid \mathcal{O}), o_i)$, in which the conditional $woe(g/g' : o_i \mid \mathcal{O})$ for each paired hypothesis g and g' is computed for each added observation o_i to the observed sequence $\mathcal{O}$ each time step. The WoE is computed using:

$$woe(g/g' : o_i \mid \mathcal{O}) = \log \frac{P(g \mid \mathcal{O}_i)}{P(g' \mid \mathcal{O}_i)} \quad (2)$$

Algorithm 1: Explanation Generation Algorithm

Input: $\mathcal{O}$, G_p, G_c, and $Posterior\ probability$ over $\mathcal{G}$
Output: Explanation list Ω of all G_p paired with G_c

for $o_i \in \mathcal{O}$ **do**
  $\Omega \leftarrow []$
  for each $g \in G_p$ **do**
    for each $g' \in G_c$ **do**
      $\omega_i \leftarrow woe(g/g' : o_i \mid \mathcal{O})$
      $\Omega \leftarrow [(g, g') = \langle \omega_i, o_i \rangle]$
    end for
  end for
end for
return Ω

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 in Algorithm 1.

Example 1. In the navigational GR scenario presented in Figure 1, a complete explanan has been generated for the observation sequence, $\mathcal{O} = \langle o_1, .., o_8 \rangle$. For the first four observations, o_1 to $o_4 \in \mathcal{O}$, the WoE would be the same for all goal hypotheses. This is because the Mirroring GR algorithm predicts them as equally likely since these observations are part of the optimal plan to achieve all three goals.

However, this will not be the case for the rest of the observation sequence. For observations o_5 to $o_7 \in \mathcal{O}$, the Mirroring GR algorithm output would be goals $g2$ and $g3$ (both equally likely), and the counterfactual goal would be $g1$. After observation, o_8, the predicted goal would be g_2, and the counterfactual goals would be $g1$ and $g3$. Table 1 presents the generated complete explanan with each new observation.

Explanation Selection

The task of explaining a GR algorithm's output in terms of the complete explanan would be tedious or even impossible, particularly in a domain where an MDP model contains hundreds of thousands of states and actions. Indeed, 'good' explanations should be selective by focusing on one or two possible causes instead of all possible causes for a decision or recommendation (Miller 2019). In the context of GR, people explain in terms of the most important observation to achieve certain goals in contrast with other alternatives (Alshehri et al. 2021). To this end, we focus on the explanation selection of answering 'Why g?' and 'Why not g'?' questions.

o_i	(g, g')	ω_i	Ω
o_5	(g_2, g_1)	0.28	$\langle 0.28, o_5 \rangle$
	(g_3, g_1)	0.28	$\langle 0.28, o_5 \rangle$
o_6	(g_2, g_1)	0.51	$\langle 0.51, o_6 \rangle$
	(g_3, g_1)	0.51	$\langle 0.51, o_6 \rangle$
o_7	(g_2, g_1)	0.69	$\langle 0.69, o_7 \rangle$
	(g_3, g_1)	0.69	$\langle 0.69, o_7 \rangle$
o_8	(g_2, g_1)	0.85	$\langle 0.85, o_8 \rangle$
	(g_2, g_3)	0.18	$\langle 0.18, o_8 \rangle$

Table 1: Complete explanan WoE for predicted and counterfactual goals after observations o_4, o_5 and o_6 in the navigational GR example depicted in Figure 1

'Why' Questions

Answering *why goal g?* questions, as in, *Why is goal g predicted as the most likely goal candidate?* relies on identifying the most important observation(s) that support the achievement of that goal. Following Alshehri et al. (2021), we call such observations *observational markers*.

Definition 4. Given a complete explanan of g, the *observational markers* (*OMs*) are the observation(s) that have the highest WoE value:

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

There may be multiple such observations, in which case we select them all.

Example 2. Let us go back to the navigational GR scenario and answer the question *Why g_2?*. From the *complete explanan* of g_2, shown in Table 1:

$$(g_2, g_1) = [\langle 0.28, o_5 \rangle, \langle 0.51, o_6 \rangle, \langle 0.69, o_7 \rangle, \langle 0.85, o_8 \rangle]$$
$$(g_2, g_3) = [\langle 0.18, o_8 \rangle]$$

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.*

'Why Not' Questions

The question of *why not g'* relies on identifying the most important observation(s) to g', which Alshehri et al. (2021) call a *counterfactual observational markers*.

Definition 5. Given a complete explanan of g', the *counterfactual observational markers* (*counterfactual OMs*) are the observation(s) that have the lowest WoE value:

$$counterfactualOM = \arg\min_{o_i \in \mathcal{O}} [(g, g') = \langle \omega_i, o_i \rangle] \quad (4)$$

Example 3. Let us go back to the navigational GR scenario and answer the question *Why not g_1 and g_3?*. From the *complete explanan* of g_1 and g_3, shown in Table 1:

$$(g_2, g_1) = [\langle 0.28, o_5 \rangle, \langle 0.51, o_6 \rangle, \langle 0.69, o_7 \rangle, \langle 0.85, o_8 \rangle]$$
$$(g_2, g_3) = [\langle 0.18, o_8 \rangle]$$

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

Counterfactual Action Pointing to the lowest WoE action is not such a useful way to understand why a counterfactual goal is not predicted. Alshehri et al. (2021) show that part of being able to answer 'why not' questions is the ability to reason about the counterfactual actions that should have happened instead of *counterfactual OM* for g' to be the predicted goal (Alshehri et al. 2021). It is called *counterfactual plan*.

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

Definition 6. Given a *counterfacual OM* o_i at state s_{t-1} for counterfactual goal g', a *counterfactual action explanation* is the action a'_t, which is the first action from the plan $\pi = (a'_t, a'_{t+1}, .., g')$ generated by solving the planning problem $\langle M, s_{t-1}, g' \rangle$, where M is the planning domain.

Example 4. Consider again the example from Figure 1. The counterfactual action a'_t would be the *move up* action from cell 23 to 14 for g_1, and the *move right* action from cell 26 to 27 for g_3 (as presented by the red arrows). Verbally, the complete explanation to 'Why not goal g_1 andg_3?' 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 it has moved up from cell 26 to cell 17, it would have moved right from cell 26 to cell 27 if the goal was* g_3.

Computational Evaluation

We evaluate the computational cost of the XGR model over eight online GR benchmark domains (Vered et al. 2018). The benchmark domains vary in the levels of complexity and size including the different number 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 cost of the planner and the explanation generation and show what effect it has on overall model performance.

Table 2 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 is spent on planning. The varying percentage increases between domains like Zeno-Travel and Kitchen emphasize the relation between the domain complexity and planning time, the higher the domain complexity, the higher influence the planner has. This highlights the impact that planner choice can have on our model performance. As the XGR approach is independent of the underlying GR model, this also shows that the proposed model would scale well with more efficient planners; e.g. domain-specific planners.

Empirical Evaluation: Human Study

Human Study 1: Ground Truth

As there is no other explainable GR approach we do not have a baseline against which to compare the explanations of our approach. We, therefore, conducted a human subject study in which participants were presented with the output of the Mirroring GR algorithm (Vered et al. 2018) and required to answer questions about its recognition process. By comparing human-generated explanations to the output of our XGR model, we aim to evaluate whether the model output is grounded on human-like explanations.

Methodology We presented participants with the Mirroring GR algorithm output over a range of problems in the Sokoban game domain, a classic warehouse puzzle game. To evaluate our hypothesis, we used the method of annotator agreement and ground truth whereby human annotation of representative features provides the ground truth for quantitative evaluation of explanation quality (Mohseni, Block, and Ragan 2018). We evaluated our XGR model by comparing it against this ground truth. Ethics approval was obtained from our institute before the study was performed.

Experiment Design We built a modified version of the Sokoban game (the interface is shown in Figure 2). Sokoban is a well-known, puzzle game in which a player moves boxes around a warehouse and delivers them to target storage locations. For our purpose, we modified the game rules by enabling the player to push more than one box simultaneously. This made predicting the target goal non-trivial.

To obtain a human explanation, i.e. 'ground truth', over the Sokoban GR task, we asked three annotators (one male, two female) recruited from the graduate student cohort at our university to annotate 15 scenarios, along with their preferred explanations. Participants were aged between 29 to 40

Figure 2: Sokoban game, scenario 5 (game version 3). The blue spot marks the initial state of the agent, and the orange spot 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 (8 observations). The predicted goal is (g_1, g_2).

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 2: Performance results of the XGR model, generating explanations for the *Mirroring* GR over eight benchmarks. Column 1 shows the mean and standard deviation run time of XGR model with the mirroring GR. Column 2 shows the average and standard deviation run time of XGR model only, without the GR algorithm. Column 3 shows the increase in run time (as a percentage) of adding the XGR to the GR. Column 4 shows how much time (as a percentage of column 3) of the XGR model was spent in counterfactual planning.

($Mean = 33$). No prior knowledge was required. Each experiment ran for approximately 60 minutes. The annotation was conducted over four stages:

1. The game instructions were introduced to the annotator with a training scenario to help them understand the task.
2. The annotator watched a partial scenario (video clip) in which a Sokoban player tried to achieve a goal. The goals were either delivering/pushing a box to a single destination cell or delivering/pushing two boxes to two different destination cells.
3. After watching the observations, annotators were given the set of predicted goals, and counterfactual goals, that were predicted by the Mirroring GR algorithm.
4. The annotators were asked to annotate the most important observation, or optionally the two most important observations, from the observation sequence that answered the two questions: 'Why goal g?' and 'Why not goal g'?', where g was the predicted goal and g' was the counterfactual goal. Participants were also required to annotate a counterfactual action for 'Why not goal g'?'. This was obtained by asking the participant to propose a *non-observed* action that they believed would signify a move to the alternative goal g'.

The data was collected over three game versions: one version required the delivery of a single box to one destination (game version 1), or two sequential destinations with interleaved plans to achieve them (game versions 2 and 3). The difference between game versions 2 and 3 pertained to the agent's ability to push multiple boxes in game 3, whereas in game 2 the agent could only push one box at a time. Each version was comprised of five different scenarios varying by complexity (15 scenarios in total). Each scenario depicted a different GR problem in which there were several competing goal hypotheses. The final destinations were not disclosed to the participants.

We combined the three annotations into a single ground truth using a majority vote. In the case of disagreements between annotations, they are often resolved by adjudicating by a fourth annotator. In our experiments, there were no disagreement instances.

We ran our model over the online GR mirroring implementation model for each of the 15 scenarios. We obtained the explanan list of XGR model to answer the *Why goal g?* question by selecting the observations with the highest WoE (OM), and the *Why not goal g'?* question by selecting the observations with the lowest WoE (*counterfactual* OM). We then compared the output explanations of our model to the ground truth obtained by human annotation. To do this, each observation in the ground truth was given its equivalent rank in the model-generated ranked explanan list.

Example 5. Considering the example from Figure 2, we obtain the complete explanan for both questions from our model. The annotated observations from the 'ground truth' is o_7 and o_2 that explain 'Why $(g1, g2)$?', and o_2 that explains 'Why not $((g3, g4)(g5, g6))$?'. We then assigned rank values to them based on the explanan list of each question, that is for why, rank values start from the highest, thus the annotated observation rank is o_2 = 7, and o_7 = 2 (o_7 rank value is 2 as it has the second highest WoE after o_8 that has rank value 1). For why not, rank values start from the lowest, thus the annotated observation rank is o_2 = 1 (o_2 is the lowest with rank value 1 as it has the lowest woe).

We then calculated the mean absolute error (MAE) to analyze how close the obtained explanation of the XGR model was to the ground truth. The MAE was calculated as the differences between each ground truth value ($a_{groundTruth}$) and XGR value (a_{XGR}) for the instance, averaged on the length of the observation sequence (n).

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

The 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, and an error of 1 means that the top-ranked observation was ranked last by the XGR model.

For evaluating the selection of a counterfactual action, as there is only one counterfactual action per plan, this is a binary agreement between ground truth and the XGR model. For each domain, we calculated the percentage of agreements using (Araujo and Born 1985):

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 3: 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.

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

Results The results of the comparison are presented in Table 3. Each row shows the MAE value calculated for each game scenario. The 'Why' and 'Why not' columns represent the MAE for our model compared to the human ground truth, and the CF(%) column represents the percentage of counterfactual action explanations that agreed with the human ground truth. We can see that for the majority of instances, the XGR model agreed completely with the ground truth obtained through human annotation. When answering *Why g?* questions the model agreed with the ground truth over 11 of the 15 scenarios (73.3%) and when answering *Why not g'?* questions the model agreed with the ground truth over 14 of the 15 scenarios (93.3%).

The CF column represents the percentage of counterfactual explanations that agreed with the human ground truth, so in this instance, higher values are better, with 100% indicating full agreement. The full agreement rate was obtained in 11 of the 15 scenarios (73.3%).

Investigating scenario 5 in game version 3, which has a relatively high MAE for the *Why goal g?* question. The scenario can be seen in Figure 2 and involves the agent delivering 2 boxes to 2 different locations, while also being able to push 2 boxes at the same time. The blue arrows represent the observation sequence, whereby the agent started at the blue circle and moved along the arrows to its current location.

In this scenario the most likely goal candidate, as predicted by the GR was delivering Box1 to g_1 and delivering Box2 to g_2, i.e. $G_p = \{(g_1, g_2)\}$. The counterfactual goal candidates were delivering the boxes to either g_3 and g_4 or g_5 and g_6, $G_c = \{(g_3, g_4), (g_5, g_6)\}$. It is important to note that in order to push both boxes simultaneously to goal (g_1, g_2), the agent would need to stand on the right of the boxes, but to push both boxes simultaneously to either (g_3, g_4) or (g_5, g_6), it needs to position itself on the left.

The XGR model's explanation to the *Why goal g?* question is the observation o_8. This can be seen as the agent aiming to position itself on the right of both boxes, confirming the supposition that the goal is (g_1, g_2). Fully considering the agent's ability to push multiple boxes, this observation constitutes the OM, the observation with the highest 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 WoE, actually making it the *counterfactual* OM 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 choose to use the same answer for both *Why goal g?* and *Why not g'?* questions can also be found in the 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 side of participants. Thus, we had a follow-up experiment where we presented the participants with the scenarios they had confusion with (Table 3, scenarios of bold values). For each scenario, we provided them with two explanation systems to answer the two questions where the first system explanation is given by our model, and the second system explanation is given by the ground truth. Then we asked them which system provide a better explanation. For the three participants, the chosen system was the first one (our model).

Human Study 2: Explainability

We conduct a second human subject experiment to evaluate the explainability of our model. We consider two hypotheses for our evaluation; 1) our model (XGR) leads to a better understanding of a GR agent; and 2) a better understanding of an agent fosters user trust in the agent.

Methodology We used the Sokoban game as the domain of the GR agent. We presented participants with the Mirroring GR algorithm output over six problems in Sokoban game domain. To evaluate hypothesis 1, we used the task prediction method (Hoffman et al. 2018). Task prediction is a proxy measure for user understanding. Participants are instructed to predict the goals of the agent. We used the *explanation satisfaction scale* by Hoffman et al. (2018) to measure the explanation's subjective quality. To evaluate hypothesis 2, we used the *trust scale* by Hoffman et al. (2018).

Experiment Design We presented six partial scenarios (video clips) of the Sokoban player trying to achieve the goal of delivering boxes to certain locations. The independent variable in this experiment is the explanation type: (1) our explanation model (XGR); and a baseline of no explanation. There is no baseline of another explanation method since there is no other method as far as we know.

The experiment has four phases. The first phase involves a collection of demographic information and training the par-

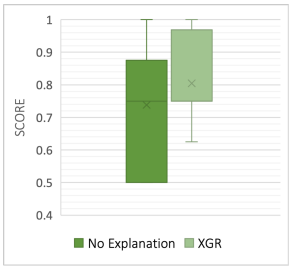

Figure 3: Task prediction scores for the two models (higher is better).

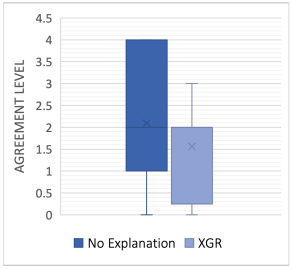

Figure 4: Behavioural trust for the two models (lower is better).

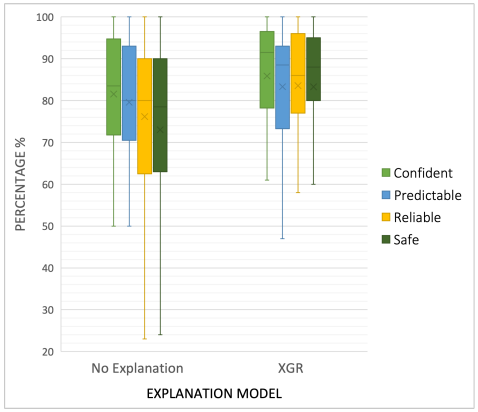

Figure 5: Perceived trust metrics for the two models on the Likert scale percentage with means represented as markers (higher indicates strongly agree).

ticipants. The participant is trained to understand the player task and how the GR system works using two video clips. In the second phase, a video clip (10 sec) is played and the GR system output is shown. Participants were asked to answer "what is the agent's predicted goal(s)?". For the baseline condition, they answered it without receiving any explanations. In the XGR condition, our model explanations for 'why' and 'why not' questions were presented. Participants completed six scenarios each. The explanations are pre-generated from our implemented algorithm and displayed on an annotated picture of the video clip's last frame. In the third phase, participants were asked to complete the trust scale. The second condition has a fourth phase of completing the explanation satisfaction scale.

We conducted the experiments on Amazon MTurk with 60 participants, allocated randomly and evenly to each condition. Each experiment ran approximately 20 minutes, and we compensated each participant with $4USD. A bonus compensation of $0.20USD was given for each correct prediction, for a total of $1.20USD. Participants were aged between 31 to 60 ($Mean = 41.4$), 23 male and 37 female; none selected non-binary or self-specified. To ensure data quality, we recruited only 'master class' workers with 98% or more approval rates.

Results We first show our results on the first hypothesis, corresponding null and alternative hypothesis are, $H_0 : S_{XGR} = S_{NoExplanation}$; and $H_1 : S_{XGR} \geq S_{NoExplanation}$ where S denotes the participants' prediction scores. Figure (3) shows the task score variance with the two models. A Welch two-sample t-test resulted in p-value of **0.05**, thus we reject H_0 and accept the alternative hypothesis H_1. These results demonstrate that the XGR model leads to a significantly better understanding of the agent's behavior than the baseline model.

Table 4 shows the average and standard deviation of the explanation satisfaction metrics on a Likert scale percentage

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

Table 4: Mean and standard deviation of explanation quality metrics for XGR model

(a higher value indicates stronger agreement). These results indicate a satisfactory level across the four metrics.

Next, we evaluate the second main hypothesis: a better understanding of an agent fosters user trust in the agent. $H_0 : T_{XGR} = T_{NoExplanation}$; and $H_1 : T_{XGR} \geq T_{NoExplanation}$ where T denotes the participants' perceived trust of the agent. Figure (5) shows the trust rate variance with the two models. We performed a Welch Two Sample t-test and obtained p-values **0.12**, **0.18**, **0.04**, and **0.02** for trust metrics *confident*, *predictable*, *reliable*, and *safe* respectively. We reject H_1 for the first two metrics and accept it for the rest. Results indicate a significant difference for *reliable*, and *safe* metrics and no difference for *confident*, *predictable*. Although the scores and trust are significantly better for our model, further investigation in a more complex domain is required.

We also evaluate behavioral trust by measuring the participants' level of agreement with the GR predictions, the difference between the GR prediction, and the participants' correct predictions (task prediction score). Clearly, the trend between the two models shown in Figure 4 is the same as for the perceived trust (Figure 5).

Conclusion

We have introduced an explanation model for GR algorithms, called eXplainable Goal Recognition (XGR). Our model generates explanations answering both 'why' and 'why not' questions pertaining to the problem of GR. Our approach builds upon the WoE concept. We computationally evaluated the performance of our system. We also conducted two human studies, which showed that our model generates explanations consistent with human labelers in over 73% of scenarios, and demonstrated that the XGR model improves people's ability to predict an agent's goal and trust in the GR algorithm. In the future, we plan to extend our work to explain noisy observation sequences, as well as conduct further evaluations to determine the source of differences between the model's explanation and human-generated explanations.

References

Akula, A. R.; Wang, K.; Liu, C.; Saba-Sadiya, S.; Lu, H.; Todorovic, S.; Chai, J.; and Zhu, S.-C. 2022. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *Iscience*, 25(1): 103581.

Albrecht, S. V.; Brewitt, C.; Wilhelm, J.; Gyevnar, B.; Eiras, F.; Dobre, M.; and Ramamoorthy, S. 2021. Interpretable Goal-based Prediction and Planning for Autonomous Driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 1043–1049. ieeexplore.ieee.org.

Alshehri, A.; Miller, T.; Vered, M.; and Alamri, H. 2021. Human centered explanation for goal recognition system. In Miller, T.; Weber, R.; Aha, D.; Magazzeni, D.; and Amir, O., eds., *IJCAI-PRICAI Workshop On Explainable Artificial Intelligence (XAI), 2020*.

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

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

Brewitt, C.; Gyevnar, B.; Garcin, S.; and Albrecht, S. V. 2021. GRIT: Fast, Interpretable, and Verifiable Goal Recognition with Learned Decision Trees for Autonomous Driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1023–1030. ieeexplore.ieee.org.

Brewitt, C.; Tamborski, M.; and Albrecht, S. V. 2022. Verifiable Goal Recognition for Autonomous Driving with Occlusions. *arXiv preprint arXiv:2206.14163*.

Cashmore, M.; Collins, A.; Krarup, B.; Krivic, S.; Magazzeni, D.; and Smith, D. 2019. Towards explainable AI planning as a service. *arXiv preprint arXiv:1908.05059*.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2018. Human-Aware Planning Revisited: A Tale of Three Models. In *IJCAI-ECAI XAI/ICAPS XAIP Workshops*.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2020. The emerging landscape of explainable ai planning and decision making. *arXiv preprint arXiv:2002.11697*.

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

Cranefield, S.; Winikoff, M.; Dignum, V.; and Dignum, F. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *IJCAI*, 178–184.

de Graaf, M. M.; Malle, B. F.; Dragan, A.; and Ziemke, T. 2018. Explainable robotic systems. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 387–388.

Fitzpatrick, G.; Lipovetzky, N.; Papasimeon, M.; Ramirez, M.; and Vered, M. 2021. Behaviour Recognition with Kinodynamic Planning Over Continuous Domains. *Frontiers in Artificial Intelligence*, 4: 717003.

Fukuchi, Y.; Osawa, M.; Yamakawa, H.; and Imai, M. 2017. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In *Proceedings of the 5th International Conference on Human Agent Interaction*, 97–101.

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.; and Albrecht, S. V. 2021. Interpretable Goal Recognition in the Presence of Occluded Factors for Autonomous Vehicles. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7044–7051. ieeexplore.ieee.org.

Heider, F. 2013. *The psychology of interpersonal relations*. Psychology Press.

Hempel, C. G. 2013. Rational action. *The American Philosophical Association Centennial Series*, 85–102.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hoffmann, J.; and Magazzeni, D. 2019. Explainable AI planning (XAIP): overview and the case of contrastive explanation. *Reasoning Web. Explainable Artificial Intelligence*, 277–282.

Hu, Y.; Xu, K.; Subagdja, B.; Tan, A.-H.; and Yin, Q. 2021. Interpretable Goal Recognition for Path Planning with ART Networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. ieeexplore.ieee.org.

Kaminka, G.; Vered, M.; and Agmon, N. 2018. Plan recognition in continuous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kashima, Y.; McKintyre, A.; and Clifford, P. 1998. The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1(3): 289–313.

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

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

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2019. Explainable reinforcement learning through a causal lens. *arXiv preprint arXiv:1905.10958*.

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

Malle, B. F.; Knobe, J.; O'Laughlin, M. J.; Pearce, G. E.; and Nelson, S. E. 2000. Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology*, 79(3): 309.

Masters, P.; and Vered, M. 2021. What's the Context? Implicit and Explicit Assumptions in Model-Based Goal Recognition. In *IJCAI*, 4516–4523.

Melis, D. A.; Kaur, H.; Daumé III, H.; Wallach, H.; and Vaughan, J. W. 2021. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 35–47.

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

Meng; Zhao; Chang; Huang; and others. 2019. Interpretable spatio-temporal attention for video action recognition. *Proc. Estonian Acad. Sci. Biol. Ecol.*

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

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

Mohseni, S.; Block, J. E.; and Ragan, E. D. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*.

Pereira, R.; Oren, N.; and Meneguzzi, F. 2017. Landmark-based heuristics for goal recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Pereira, R. F.; Vered, M.; Meneguzzi, F. R.; and Ramírez, M. 2019. Online probabilistic goal recognition over nominal models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, China.*

Ramírez, M.; and Geffner, H. 2010. Probabilistic plan recognition using off-the-shelf classical planners. In *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.

Shahrdar, S.; Menezes, L.; and Nojoumian, M. 2018. A survey on trust in autonomous systems. In *Science and Information Conference*, 368–386. Springer.

Shvo, M.; and McIlraith, S. A. 2020. Active goal recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9957–9966.

Sohrabi, S.; Riabov, A. V.; and Udrea, O. 2016. Plan Recognition as Planning Revisited. In *IJCAI*, 3258–3264. New York, NY.

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

Vered, M.; Howe, P.; Miller, T.; Sonenberg, L.; and Velloso, E. 2020. Demand-driven transparency for monitoring intelligent agents. *IEEE Transactions on Human-Machine Systems*, 50(3): 264–275.

Vered, M.; and Kaminka, G. A. 2017. Heuristic online goal recognition in continuous domains. *arXiv preprint arXiv:1709.09839*.

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

Vered, M.; Pereira, R. F.; Magnaguagno, M. C.; Kaminka, G. A.; and Meneguzzi, F. 2018. Towards Online Goal Recognition Combining Goal Mirroring and Landmarks. In *AAMAS*, 2112–2114.

Winikoff, M.; Dignum, V.; and Dignum, F. 2018. Why Bad Coffee? Explaining Agent Plans with Valuings. In *International Conference on Computer Safety, Reliability, and Security*, 521–534. Springer.

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