A Hypothesis-Driven Approach to Explainable Goal Recognition

Abeer Alshehri
University of Melbourne
Melbourne, Australia
King Khalid University
Abha, Saudi Arabia
adshehre@kku.edu.sa

Hissah Alotaibi
Jazan University
Jazan, Saudi Arabia
halotaibi@jazanu.edu.sa

Tim Miller
University of Queensland
Brisbane, Australia
timothy.miller@uq.edu.au

Mor Vered
Monash University
Melbourne, Australia
mor.vered@monash.edu

ABSTRACT

In this paper, we introduce an explainable goal-recognition (XGR) approach for decision support that instantiates the evaluative AI paradigm. Current explainable AI (XAI) approaches focus on providing recommendations and justifying those recommendations. However, a shift toward evaluative AI has been proposed, focusing on generating evidence to support or refute human judgments and explaining trade-offs among hypotheses, rather than merely justifying AI recommendations. We introduce such a method for goal recognition tasks by leveraging the Weight of Evidence (WoE) framework. Through a human study in a maritime surveillance task, we demonstrate that our model improves decision accuracy, efficiency, and reliance in complex scenarios, outperforming two baseline models and demonstrating its potential in real-world decision-making.

KEYWORDS

Explainable AI; Evaluative AI; Goal recognition; Planning

ACM Reference Format:
Abeer Alshehri, Hissah Alotaibi, Tim Miller, and Mor Vered. 2025. A Hypothesis-Driven Approach to Explainable Goal Recognition. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

The collaboration between humans and Artificial Intelligence (AI) is often motivated by the understanding that their combined strengths can accomplish more than either could alone. AI-assisted decision making, where AI provides recommendations to human decision makers, is becoming increasingly common in fields such as medical diagnostics and criminal justice [10, 12, 26, 29]. However, since AI systems are not perfect, human decision makers must judge when to trust them. This can lead to *overtrust*, users rely on AI too much, or *undertrust*, users are overly skeptical despite AI's capabilities [14, 42]. Encouraging appropriate reliance is essential for effective collaboration, and recent research has focused on methods to better align users' trust with the actual performance of AI systems.

Explainability is essential to encouraging trust and appropriate reliance on AI systems. It serves as a transparency mechanism, providing insight into how a system operates, enabling users to assess the accuracy of its outputs and identify potential errors [11, 40, 43]. Current approaches to explainable AI, often termed the *recommendation-driven* decision support model [35], focus on delivering the "best" recommendation and explaining why it was chosen, even when the system is uncertain about its correctness. While this approach is intended to build trust, empirical evidence suggests it may have the opposite effect. Several studies have demonstrated that this method can actually impede appropriate reliance, as users may find the explanations unconvincing or incomplete [3, 7, 27, 49]. This limitation reduces users' ability to critically engage with AI recommendations, diminishing the effectiveness of *recommendation-driven* decision support in promoting trust [35].

To address these limitations, there is a growing call for a paradigm shift toward *evaluative AI* [35], which provides a more robust framework to enable users to critically engage with AI systems. This approach moves away from *recommendation-driven* decision support and focuses on *hypothesis-driven* models, where evidence is generated for evaluative human judgments by providing evidence that supports or refutes human decisions. Instead of justifying AI recommendations, it explains the trade-offs among different hypotheses, improving trust calibration and mitigating concerns about over-reliance or under-reliance by not steering decision-makers toward specific choices [28].

Goal recognition, a key aspect of AI decision-making, involves inferring the goals or intentions behind observed actions [31]. It is essential in fields such as autonomous systems and human-robot interaction. Recent advances in goal recognition techniques aim to make these systems more efficient and interpretable, thereby enhancing decision-making capabilities while increasing user trust and satisfaction [4, 24, 51]. In this context, Alshehri et al. [1] introduced the eXplainable Goal Recognition (XGR) framework. This model, based on the concept of Weight of Evidence [18], helps decision makers understand the reasoning behind predictions from goal recognition systems.

Building on this foundation, we extend our work by introducing a *hypothesis-driven* XGR model, which leverages the Weight of Evidence (WoE) framework to improve decision-making in goal

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

recognition tasks. In this model, WoE is used to generate sets of supporting and refuting evidence for each goal hypothesis, with the most weighted evidence selected. This approach aligns with the evaluative AI paradigm [35], offering deeper insights and enabling more informed decisions. We evaluated the model in a human study involving 275 participants in the maritime surveillance domain. The empirical results demonstrate that the hypothesis-driven XGR model significantly enhances the accuracy, efficiency, and reliance of decision making in complex high-stakes scenarios, outperforming the two baseline models tested: *Soft_XGR* and *Hard_XGR*, which represent soft and hard variations of the state-of-the-art XGR approach introduced in [1].

2 RELATED WORK AND BACKGROUND

2.1 Goal Recognition (GR)

Goal recognition (GR), often situated among *Plan, Activity and Intention Recognition* [45], is the process of determining an agent's goal through a sequence of observations of its interactions within a given environment. It has been widely researched in the context of smart homes [22], daily living assistance [19], UAV detection [15], robotics [8], autonomous vehicles [21], and much more. While traditional approaches to plan recognition have employed the use of plan libraries to compare observations against, recent approaches adopt the use of planning techniques within the recognition process, labeled as Plan Recognition as Planning (PRP) [33, 41].

We follow the definition of GR established by Shvo and McIlraith [44] which accepts as input a set of goal hypotheses assuming that the intended goal(s), $g \in G$, is among them.

Definition 2.1. A goal recognition task is a tuple composed of $\langle \Xi, I, \mathcal{G}, O \rangle$, where Ξ is a domain definition, I is an initial state, $\mathcal{G}$ is a set of goal hypotheses and $mathcalO = \langle o_1, o_2, ..., o_n \rangle$ is a sequence of observations.

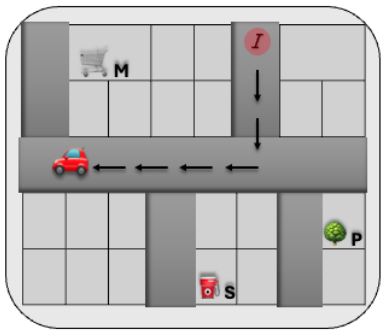


Figure 1: Navigational domain example: (I) represents the car's initial state, with the goal set $\mathcal{G}$ consisting of the Supermarket (M), Gas Station (S), and Park (P). The black arrows indicate the observation sequence $O = \langle o_1, \ldots, o_6 \rangle$.

Example 2.2. Figure 1 illustrates a goal recognition (GR) problem where a navigation system is monitoring a car as it drives through a city. The system is tasked with predicting the intended destination of the car, defined by the set of goal hypotheses $\mathcal{G} =$ {Supermarket, Gas Station, Park}. The car moves down the street from the initial state (labeled I) and takes a right turn onto Main Street. The navigation system must evaluate the sequence of observations $O = \langle o_1, \ldots, o_6 \rangle$ (represented as black arrows) to determine the most probable destination for the car.

2.1.1 Mirroring GR Algorithm. The mirroring algorithm [25, 38, 48] is a popular, PRP, GR algorithm. Mirroring employs a planner to generate optimal plans that progress from an initial state I to each goal, $g_j \in \mathcal{G}$. Then for each incoming observation, $o_i \in O$, Mirroring creates a new plan that concatenates a *prefix* (the observations seen so far) and a *suffix* (generated by another planner call which plans from the most recent observation to each of the goals) and compares this new plan against the previously generated optimal plans. The algorithm then yields a likelihood distribution, represented as posterior probabilities $P(g_j \mid O)$, for each $g_j \in \mathcal{G}$, by evaluating which of the generated plans, incorporating observations O, aligns the most closely with the optimal plan.

Referring back to the example in 2.2, before the car makes the right turn, the mirroring GR algorithm would rank all three goals as equally likely, as the first two actions are on the optimal path to each goal. However, once the car turns right, the GR algorithm would consider the probability of heading to the park as less likely, since turning right is a suboptimal action for that plan. After the car passes the turn-off for the gas station, the mirroring GR algorithm would conclude that the most likely destination for the car is the supermarket, as the observed sequence of movements aligns with the optimal plan to reach this goal and not with any other goal.

2.2 eXplainable Goal Recognition XGR

Alshehri et al. [1] proposes an eXplainable Goal Recognition (XGR) model rooted in the concept of Weight of Evidence (WoE) from information theory [18, 32]. The weight of evidence is a statistical measure that is used to quantify the influence of variables on prediction models. It is defined in terms of log-odds, which assess the strength of evidence e in favour of a hypothesis h versus an alternative hypothesis h', given additional information c [18]. Assuming uniform prior probabilities, WoE is expressed as:

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

The XGR model uses WoE values to generate counterfactual explanations for GR outputs. It evaluates how much an observation supports one goal hypothesis over another by assessing the strength of observed evidence for the predicted goal against counterfactual goals. To address the questions "Why?" and "Why not?", the model displays the posterior probabilities of hypotheses provided by the GR system, reflecting the certainty level associated with each prediction. The answers are derived through key observations, observational markers and counterfactual observational markers. Observational markers indicate the observation with the highest WoE, representing the strongest evidence for the predicted goal, while counterfactual markers highlight the observation with the lowest WoE, explaining why alternative hypotheses were not predicted. For example, in case 2.2, the answer to "Why is the supermarket the most likely destination?" would be: "Because the car moved forward at (o_6)," as this observation has the highest WoE value.

However, while these counterfactual explanations can help users understand why an alternative hypothesis was not chosen, they fall short in addressing a key limitation: they do not provide reasoning or evidence to determine whether the alternative might actually be better or more valid. The current model focuses primarily on refuting alternatives but does not fully support a critical comparison of goals that might enable users to judge whether another option should have been selected.

2.3 Evaluative AI and Decision Making

Traditional XAI methods focus on providing users with understandable explanations of AI-generated recommendations to increase trust and transparency [20]. This involves techniques such as highlighting important features, visualizing decision paths, or providing natural language explanations that justify the recommendation [52]. However, recent studies suggest that these recommendation-driven approaches have a limited impact on decision making, as users often fail to fully engage with explainability tools in a way that enhances their decision-making [17, 37, 39, 47]. This has prompted calls for XAI models that go beyond mere explanation and foster critical engagement with the decision space.

Evaluative AI emerges as a new conceptual framework [35] that shifts the AI's role from a decision maker to a thought partner, encouraging users to explore evidence supporting or refuting their hypotheses interactively. By presenting supporting and opposing evidence, this hypothesis-driven approach promotes better understanding, decision accuracy, and a more appropriate reliance on AI.

Recent work [28] has implemented evaluative AI using the WoE framework, generating supporting and opposing evidence for a given hypothesis. A human behavioral experiment demonstrated that this approach improved the accuracy of decisions and led to a more appropriate reliance on AI. Other researchers have followed a similar approach, finding that it has the potential for enhancing trust calibration in human-AI collaboration [36], leading to more effective AI decision support [53], and fostering more informed and responsible decision-making [9]. By framing XAI in this way, users are less likely to fixate on a single output or become overly reliant on the AI's initial assessment. Instead, they are prompted to think critically about the problem, consider alternative perspectives, and use AI as a tool to explore the solution space more thoroughly.

3 HYPOTHESIS-DRIVEN XGR MODEL

We model Hypothesis-Driven XGR by extending Melis et al.'s [32] WoE framework. Our model uses WoE to generate strong evidence supporting and refuting each goal hypothesis and allowing decision-makers to access this evidence precisely when needed.

In our model, instead of contrasting h with an alternative hypothesis h' (as in Equation 1), the WoE is defined for evidence e, hypothesis h, and its logical complement $\overline{h}$, which represents all other hypotheses. Assuming uniform priors[1], the WoE of e in favor of h, and conditioned on additional information c is:

$$\text{woe}(h : e \mid c) = \log\left(\frac{P(h \mid e, c)}{1 - P(h \mid e, c)} \cdot \frac{1 - P(h)}{P(h)}\right) \tag{2}$$

[1]Formula derivation is in the supplementary material

Algorithm 1 Evidence Generation Algorithm

Input: O_i, o_i, $\mathcal{G}$, and posterior probability over $\mathcal{G}$
Output: Evidence list Ω for all $g \in \mathcal{G}$

$\Omega \leftarrow [\,]$ {Initialize evidence list}
$N \leftarrow |\mathcal{G}|$ {Number of hypotheses}
for $o_i \in O_i$ **do**
  for $g \in \mathcal{G}$ **do**
    $P(g) \leftarrow \frac{1}{N}$ {Set uniform prior probability}
    $\omega_i \leftarrow woe(g : o_i \mid O_i)$ {Compute Weight of Evidence (WoE)}
    $\Omega \leftarrow \Omega \cup \{(g, \langle \omega_i, o_i \rangle)\}$ {Add evidence to list}
  end for
end for
return Ω

If $\text{woe}(h : e \mid c) > 0$, this indicates that the evidence supports hypothesis h. Conversely, if $\text{woe}(h : e \mid c) < 0$, it suggests that the evidence refutes hypothesis h. Additionally, a value of $\text{woe}(h : e \mid c) = 0$ signifies that the evidence neither supports nor refutes hypothesis h.

Our Hypothesis-Driven XGR model accepts four inputs, which any GR model can provide; (1) **An observed sequence** O_i, representing the set of evidence 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 evidence being considered; (3) **A set of possible goals,** $\mathcal{G}$, where each $g \in \mathcal{G}$ represents a hypothesis; and (4) **Posterior probabilities** $P(g \mid O_i)$ for each goal $g \in \mathcal{G}$, indicating how likely each goal is given the evidence up to and including o_i.

In this framework, the goals serve as hypotheses, the observations represent the evidence, and the current observation is the specific piece of evidence being weighed. The model computes the WoE incrementally for each goal hypothesis as new observations in the sequence O are processed.

3.1 Evidence Generation

Referring to Equation 2, we substitute the hypothesis h with a goal g, the evidence e with the observation $o_i \in O_i$, the additional information c with the observed sequence O_i up to the observation o_i, and the posterior probabilities as $P(g \mid O_i)$, in which $\mathcal{G}$. A complete explanation is defined as follows.

Definition 3.1. A *complete evidence* for a goal g is a list of pairs $(woe(g : o_i \mid O_i), o_i)$, in which the conditional weight of evidence $woe(g : o_i \mid O_i)$ for each hypothesis $g \in \mathcal{G}$ is computed for each added observation o_i to the observed sequence O_i. The WoE is computed as follows:

$$\text{woe}(g : o_i \mid O_i) = \log\left(\frac{P(g \mid O_i)}{1 - P(g \mid O_i)} \cdot \frac{1 - P(g)}{P(g)}\right) \tag{3}$$

where $O_i = \{o_1, o_2, \ldots, o_i\}$ represents the sequence of all observations up to and including the current observation o_i.

Informally, a complete evidence set for a goal g consists of the full set of computed WoE scores for each observation. An algorithm to extract these scores is presented in Algorithm 1.

If we look back to example 2.2 (Figure 1), the weight of evidence would be zero for observations o_1 and o_2 since the GR algorithm predicts all hypotheses as equally likely. After observing the car turning right, o_3 and o_4, the supermarket and gas station emerge as

the most probable goal destinations. Then, following observation o_5 and o_6 when the car passes the turn-off, the supermarket becomes the single most probable goal. Table 1 presents the complete explanations, (*woe, observation*) pair, for each goal g.

Table 1: Complete Set of Evidence for Each Goal

Goal g	Complete Evidence Set
Park	$\langle 0, o_1\rangle, \langle 0, o_2\rangle, \langle -0.33, o_3\rangle, \langle -0.56, o_4\rangle, \langle -0.67, o_5\rangle, \langle -0.76, o_6\rangle$
Gas Station	$\langle 0, o_1\rangle, \langle 0, o_2\rangle, \langle 0.22, o_3\rangle, \langle 0.26, o_4\rangle, \langle 0.09, o_5\rangle, \langle -0.11, o_6\rangle$
Supermarket	$\langle 0, o_1\rangle, \langle 0, o_2\rangle, \langle 0.22, o_3\rangle, \langle 0.26, o_4\rangle, \langle 0.51, o_5\rangle, \langle 0.64, o_6\rangle$

3.2 Evidence Selection

Effective explanations should be selective, concentrating on one or two potential causes rather than attempting to address all possible causes for a decision or recommendation [34]. People frequently reference the observational marker when formulating explanations for their decisions, as it represents the most critical observed evidence for the goal hypothesis [2]. We selected the evidence that **supports** the hypothesis by following the definition of the *observational marker* established by Alshehri et al. [1].

Definition 3.2 (Supporting Marker). Given the complete evidence set of g, the *supporting markers* (*SMs*) are the observed actions with the highest Weight of Evidence (WoE) values:

$$SM = \arg\max_{o_i \in O_i} \omega_i,$$

where ω_i is the WoE value associated with the observed action o_i.

Our approach to identifying refuting evidence highlights the most critical observed evidence in the sequence. In sequential tasks, such as GR tasks, the evidence that refutes the goal hypothesis accumulates over time, meaning that as an agent moves further from the goal, the Weight of Evidence (WoE) decreases. However, the lowest WoE is not necessarily the best refuting evidence; for example, in the navigational example in Figure 1, the lowest WoE against *Park* will be the most recent action in the sequence. The observation with the lowest WoE to explain counterfactual goal hypotheses (g') was introduced as the *counterfactual observational marker* [1], which is relevant because the complete explanation list contains only positive WoE values generated for the goal pair (g, g'). In the *hypothesis-driven* context, we incorporate both positive and negative WoE values. Therefore, we define the evidence that best **refutes** the hypothesis as that which emphasizes the largest shift in WoE—the biggest difference—defined as:

Definition 3.3 (Refuting Marker). Given the complete evidence set of g, the *refuting markers* (*RMs*) are the observed actions that result in the largest decrease in WoE values. The change in WoE for an observed action o_i is defined as:

$$\Delta\omega_i = \begin{cases} \omega_{i+1} - \omega_i & \text{if } \omega_{i+1} < \omega_i, \\ 0 & \text{otherwise.} \end{cases}$$

The *refuting markers* are then identified as:

$$RM = \arg\min_{o_i \in O_i} \Delta\omega_i,$$

where $\Delta\omega_i$ quantifies the decrease in WoE from ω_i to ω_{i+1} for the observed action o_i. If $\omega_{i+1} \geq \omega_i$, $\Delta\omega_i$ is set to 0, effectively ignoring cases where the WoE does not decrease.

From the evidence set associated with each goal, we identify key pieces of evidence that either strongly support or refute the goal. Referring again to Example 2.2, the evidence set is generated for each goal g (Table 1). According to Definitions 3.2 and 3.3, the selected evidence is as follows:

- For g (**Park**), $\langle -0.33, o_3\rangle$ refutes g.
- For g (**Gas Station**), $\langle 0.26, o_4\rangle$ supports g, while $\langle -0.11, o_6\rangle$ refutes it.
- For g (**Supermarket**), $\langle 0.64, o_6\rangle$ supports g.

4 EMPIRICAL EVALUATION: HUMAN STUDY

We conducted a human study to evaluate our model in the context of AI-assisted decision-making, testing the following hypotheses compared to two baseline models:

(1) Hypothesis-driven XGR improves decision accuracy.
(2) Hypothesis-driven XGR enhances task efficiency by reducing decision completion time.
(3) Hypothesis-driven XGR promotes more appropriate reliance on AI-assisted decision-making.
(4) Hypothesis-driven XGR increases user trust in AI-assisted decision-making.
(5) Hypothesis-driven XGR delivers subjectively better explanations, leading to greater explanation satisfaction.

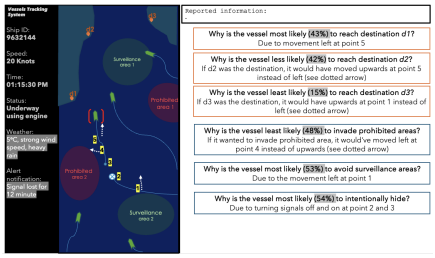
4.1 Experiment Design and Methodology

Task Setup. In this study, we focused on the illegal vessel detection domain [13], a challenging maritime environment where participants were tasked with identifying potential illegal activities. In this context, vessels may invade prohibited areas, deliberately avoid surveillance zones, or conceal their illegal operations by turning off their signals. These challenges make it an ideal setting for evaluating AI-assisted decision-making.

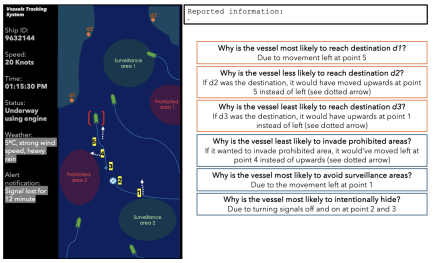
For our AI-assisted system, we used the Mirroring GR algorithm to generate goal hypotheses and provide explanations. The approach, however, is generalizable to any goal recognizer. The Mirroring GR algorithm's success and failure rates were evenly split (50% each) across tasks, ensuring participants encountered realistic conditions with varying algorithmic accuracy.

Participants encountered six scenarios in random order to mitigate ordering effects. The first two were simple, relying on a single information source, while the remaining four were complex, requiring multiple sources for realistic decision-making.

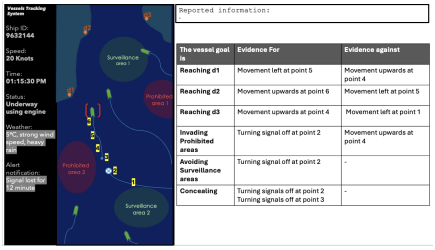
Figure 2 depicts an example scenario in which a detected vessel (red brackets) is en route to one of three possible destinations (d1, d2, or d3). Starting from the initial state (illustrated by a blue line), multiple goal hypotheses need to be considered, such as whether the vessel is invading prohibited areas, avoiding surveillance, concealing its activities, or its intended destination. The Mirroring GR algorithm provides posterior probabilities over these goal hypotheses based on the vessel's observed behavior (also represented by the blue line). Key model variables include the vessel's location,



(a) Soft_XGR



(b) Hard_XGR



(c) Hypothesis_driven_XGR

Figure 2: Example scenario in maritime surveillance across the three conditions.

the positions of prohibited and surveillance areas, potential destinations, and signal status. Participants were asked to make two different but related decisions: first, to assess the likelihood of the vessel's destination, and second, to determine whether Coast Guard intervention was necessary based on that assessment. Intervention is triggered if the probability of illegal fishing is assessed to be above a predetermined threshold, typically 50%.

Conditions. We conducted a between-subjects study with participants randomly assigned to one of three AI-assisted decision-making conditions:

- *Soft_XGR*: Participants used the XGR system defined by Alshehri et al. [1] (Figure 2 (a)), which included posterior probabilities and explanations from the Mirroring GR system. The term 'soft' refers to the concept of a 'soft' machine learning classifier [50].
- *Hard_XGR*: Participants used the same XGR system (Figure 2 (b)), but only the goal hypothesis was provided without probability. This condition is termed 'hard' based on the concept of a hard machine learning classifier [50].
- *Hypothesis_driven_XGR*: Participants made decisions based on our proposed model (Figure 2 (c)).

The inclusion of the Hard_XGR approach allows us to assess whether differences in over- or under-reliance between the original XGR model and our hypothesis-driven model stem from the probability distributions.

Procedure. The experiment was structured into four phases:

(1) **Phase 1: Demographic Data Collection.** Participants provided demographic information.
(2) **Phase 2: Training.** Participants were trained using two scenarios to familiarize themselves with the task and AI-assisted decision-making tools.
(3) **Phase 3: Task Execution.** Participants viewed a static image simulating a vessel-tracking system displaying six scenarios of a vessel navigating toward one of three seaports, accompanied by AI outputs and pre-generated explanations (Figure 2). They predicted the vessel's destination and assessed potential illegal activities requiring Coast Guard intervention.
(4) **Phase 4: Trust and Explanation Satisfaction Assessment.** Participants rated their trust in the AI outputs and completed a scale measuring satisfaction with the provided explanations.

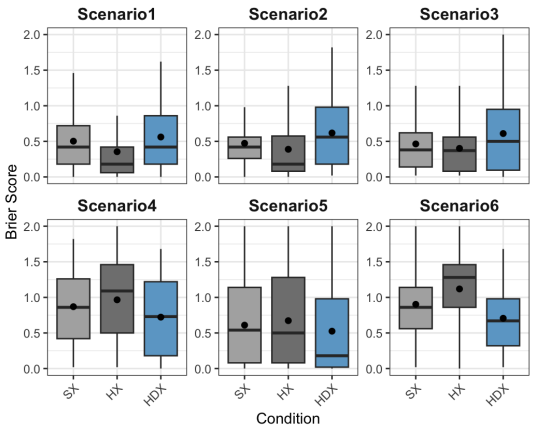
Participants. We conducted a power analysis using Cohen's F, assuming a small effect size (f = 0.20), 0.80 power, and a 0.05 significance level, yielding a target sample of 246. We recruited 283 participants via Prolific, randomly assigning them equally to the three conditions. To ensure data quality, participants were required to be native English speakers from the US, UK, or Australia, with a 99% approval rating and at least 1,000 prior submissions. After excluding inattentive respondents, we retained 275 valid responses (*Soft_XGR*: 89, *Hard_XGR*: 96, *Hypothesis_driven_XGR*: 90). The sample comprised 157 males, 117 females, and 1 self-identified, aged between 25 and 55, with a mean age of 37. Participants received $8.00 USD plus up to $3.00 USD in performance bonuses.

Metrics. We evaluated our first hypothesis using the Brier score function, which measures the accuracy of predictive probabilities for binary and multiclass outcomes. The Brier score ranges from 0 (best performance) to 1 (poorest performance) and is computed as the mean squared difference between the predicted probabilities and the actual outcomes (ground truth) [5]:

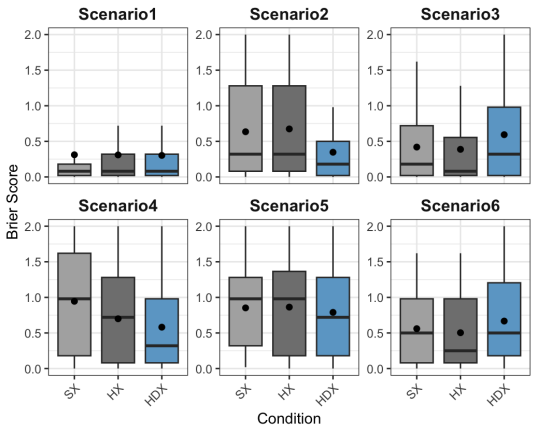
$$\text{Brier Score} = \sum_{i=1}^{c} (p_i - y_i)^2 \tag{4}$$

where c is the number of classes, p is the predicted probability, and y is the ground-truth label vector $y = (y_1, \ldots, y_c)$, with $y_i = 1$ for the true class and 0 otherwise. The Brier score, used to evaluate participants' decisions across six scenarios, indicates accuracy with lower scores representing better performance. It rewards correct decisions made with high certainty and penalizes incorrect ones, effectively mitigating the influence of random guessing or uncertainty in responses.

We addressed task efficiency in our second hypothesis by applying a logarithmic transformation to the time spent on all scenarios for both tasks, recorded in seconds. This approach normalizes the distribution of response times and reduces the influence of outliers,



(a) Predicting the vessel's destination task.



(b) Predicting the need to dispatch the Coast Guard task .

Figure 3: Brier scores across the six scenarios (lower is better). SX: *Soft_XGR*, HX: *Hard_XGR*, HDX: *Hypothesis_driven_XGR*.

Table 2: Pairwise differences of Brier score for the six scenarios. Hypothesis: *Hypothesis_driven_XGR*

Task	Scenario	Hypothesis - Soft_XGR		Hypothesis - Hard_XGR		Soft_XGR - Hard_XGR	
		Z-Value	p-adj	Z-Value	p-adj	Z-Value	p-adj
Vessel Destination Task	Scenario 1	0.21	1.00	3.61	**< 0.001**	3.38	**< 0.001**
	Scenario 2	1.42	0.23	4.52	**< 0.001**	3.07	**< 0.001**
	Scenario 3	1.37	0.25	2.48	**0.02**	1.08	0.42
	Scenario 4	-1.82	0.10	-3.04	**< 0.001**	-1.19	0.35
	Scenario 5	-1.51	0.20	-2.20	**0.04**	-0.65	0.77
	Scenario 6	-2.37	**0.03**	-5.45	**< 0.001**	-3.02	**< 0.001**
Dispatching the Coast Guard Task	Scenario 1	1.46	0.22	0.14	1.00	-1.34	0.27
	Scenario 2	-3.80	**< 0.001**	-4.27	**< 0.001**	-0.40	1.00
	Scenario 3	1.44	0.23	2.28	**0.03**	0.81	0.63
	Scenario 4	-3.53	**< 0.001**	-1.10	0.41	2.49	**0.02**
	Scenario 5	-0.90	0.56	-0.94	0.52	-0.02	1.00
	Scenario 6	1.22	0.33	2.31	**0.03**	1.06	0.44

which can skew statistical analyses and lead to misleading interpretations of results [46]. For the third hypothesis, we measured the appropriateness of reliance on AI-assisted decision-making using the following metrics [30]:

$$\text{Overreliance} = \frac{\text{Incorrect human decisions with incorrect AI}}{\text{Total incorrect AI predictions}} \quad (5)$$

$$\text{Underreliance} = \frac{\text{Incorrect human decisions with correct AI}}{\text{Total correct AI predictions}} \quad (6)$$

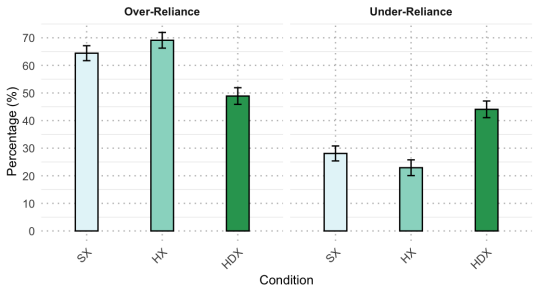
In the third condition, where participants were not directly shown the AI system's output but were instead provided with evidence, overreliance on the tool can still occur. This is because the evidence reveals aspects of the underlying decision-making process, which can implicitly guide participants toward accepting the AI system's output.

To test our fourth hypothesis, we used the Trust Scale from Hoffman et al. [23], where participants rated trust across four metrics on a 5-point Likert scale (0 = Strongly Disagree, 100 = Strongly Agree). For the final hypothesis, we assessed explanation quality using the Explanation Satisfaction Scale from Hoffman et al., with participants rating four metrics on the same 5-point scale.

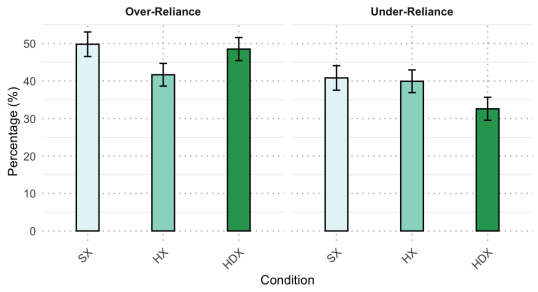
Analysis Method. For the analysis, we applied non-parametric methods, as the data did not meet the assumption of normality. The Kruskal-Wallis test assessed group differences, followed by Dunn's test with Bonferroni correction for pairwise comparisons. Additionally, we calculated the logarithmic percentage change in completion time to measure the magnitude of the difference after transformation [16]:

$$\text{Log Percentage Change} = (e^{\Delta \log} - 1) \times 100 \quad (7)$$

Where Δ log is the difference between the log-transformed means of two conditions.



(a) Predicting the vessel's destination.



(b) Predicting the need to dispatch the Coast Guard.

Figure 4: Reliance results for two tasks (lower is better). SX: *Soft_XGR*, HX: *Hard_XGR*, HDX: *Hypothesis_driven_XGR*.

5 RESULTS

5.1 Decision Accuracy

We present results on the hypothesis that *Hypothesis_driven_XGR* leads to better decisions. Brier scores across scenarios (Figures 3a and 3b) show varying effectiveness depending on scenario complexity.

In the task of predicting the vessel's destination, *Hard_XGR* consistently outperformed other conditions in scenarios 1 and 2, which were classified as simple. This is evidenced by the significant p-values presented in Table 2. In contrast, in the more challenging scenarios (3 to 6), *Hypothesis_driven_XGR* exhibited superior performance in scenarios 4, 5, and 6 compared to the baselines (Table 2). Notably, although scenario 3 was designed to be difficult, participants seemed to find it relatively easy, performing better with *Hard_XGR* than with our model (Table 2). These findings indicate that while *Hard_XGR* is more effective in simple scenarios, *Hypothesis_driven_XGR* tends to excel in more complex, real-world situations.

For the task of predicting the need to send the Coast Guard, the results are less conclusive. While *Hypothesis_driven_XGR* performs well in scenarios 2 and 4, it does not consistently outperform the other conditions in all hard scenarios (Table 2). This inconsistency may suggest that the evidence provided in certain cases is either insufficient or overly complex for participants to fully grasp within the context of the task. Further refinement of the model's presentation of evidence could enhance user comprehension.

5.2 Task Efficiency in Decision-Making

For our second hypothesis, which suggests that our model enhances overall efficiency in task completion time, results showed a significant difference between conditions ($p > 0.001$). *Hypothesis_driven_XGR* resulted in significantly faster task completion than *Soft_XGR*, with no significant difference observed between *Hypothesis_driven_XGR* and *Hard_XGR* (Table 3).

To better understand the practical significance of these findings, we examined the log percentage changes in time spent across conditions. The results showed a 15% increase in time spent from the *Hypothesis_driven_XGR* condition to the *Soft_XGR* condition, suggesting that participants took significantly longer to complete tasks with *Soft_XGR* than with *Hypothesis_driven_XGR*. In contrast, the time difference between *Hypothesis_driven_XGR* and *Hard_XGR* was smaller, with only a 5% increase in time spent. This suggests that both conditions are similarly effective in facilitating decision-making, enabling participants to complete tasks with comparable efficiency. This indicates that participants may not fully evaluate all the options the *Hypothesis_driven_XGR* offers, relying instead on the evidence to validate their thinking, resulting in faster decision-making. Conversely, when interacting with *Soft_XGR*, they appear to expend more cognitive effort trying to reconcile their thoughts when they disagree with the model's output.

Table 3: Pairwise differences of completion time.

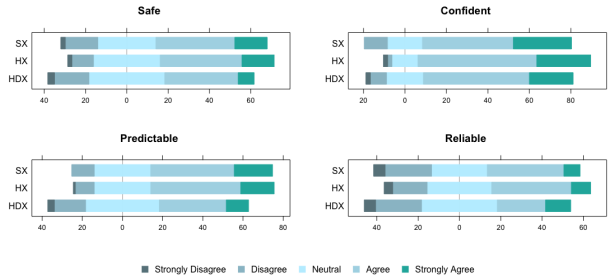
Comparison	Z-value	adj-p
Hypothesis - Soft_XGR	-3.26	< **0.001**
Hypothesis - Hard_XGR	-1.24	0.32
Soft_XGR - Hard_XGR	2.07	0.06

5.3 Appropriate Reliance on AI-Assisted Decision Making

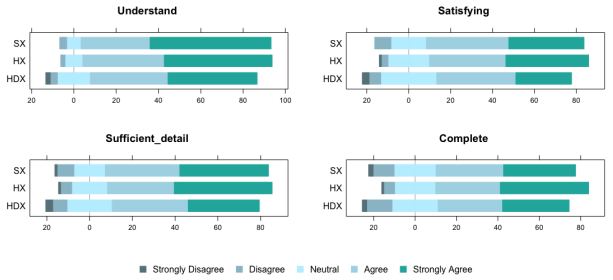
We evaluated our third hypothesis: *Hypothesis_driven_XGR promotes more appropriate reliance on AI-assisted decision-making*. Participants' reliance was assessed using metrics designed to capture both over-reliance and under-reliance on the AI system.

In the task of predicting the vessel's destination (see Figure 4a), the results for over-reliance showed significant differences across the three conditions ($p = 0.03$). Post-hoc pairwise comparisons revealed that *Hypothesis_driven_XGR* significantly reduced over-reliance compared to both *Soft_XGR* ($p = 0.02$) and *Hard_XGR* ($p < 0.001$), suggesting that *Hypothesis_driven_XGR* was more effective at mitigating over-reliance. This reduction in over-reliance aligns with previous research on cognitive forcing strategies [6, 17], where individuals are prompted to make decisions before seeing a recommendation. Such strategies encourage users to critically engage with AI outputs, thereby reducing over-reliance.

The results for under-reliance also showed significant differences across the three conditions ($p < 0.001$). Post-hoc pairwise



(a) Perceived trust metrics.



(b) Explanation quality metrics.

Figure 5: Likert scale results. The X-axis represents each Likert category's total counts of responses, adjusted to have 0 as the midpoint. SX: *Soft_XGR*, HX: *Hard_XGR*, HDX: *Hypothesis_driven_XGR*

comparisons indicated that *Hard_XGR* ($p < 0.001$) and *Soft_XGR* ($p < 0.001$) significantly reduced under-reliance compared to *Hypothesis_driven_XGR*. This increase in under-reliance may stem from the intuitive nature of predicting the vessel's destination. In such tasks, users may feel they already possess sufficient understanding to make decisions, and without explicit recommendations or predictions from *Hypothesis_driven_XGR*, they tend to rely more on their judgment and less on the evidence provided by the AI.

In the task of predicting the need to send the Coast Guard (see Figure 4b), the over-reliance results revealed that although *Hard_XGR* appeared to decrease over-reliance, there was no statistically significant difference between conditions ($p = 0.02$). For under-reliance, results showed a marginally significant difference, with a p-value of 0.11. Post-hoc pairwise comparisons revealed that *Hypothesis_driven_XGR* reduced under-reliance in this task compared to *Soft_XGR* and *Hard_XGR*, with p-values of 0.08 and 0.12, respectively. Since this task is more complex than predicting the vessel's destination, the model likely decreases under-reliance by actively engaging users with evidence that supports or refutes their hypotheses.

5.4 User Trust

To test our fourth hypothesis—that *hypothesis-driven* XGR increases user trust—we evaluated the self-reported trust scale results. We obtained p-values of 0.27, 0.04, 0.62, and 0.30 for the trust metrics *confident*, *predictable*, *reliable*, and *safe*, respectively. These results indicate significant differences in participants' perceived trust in the AI-assisted decision-making model only for the predictability metric (see Figure 5a for Likert scale results). Post-hoc pairwise comparisons showed that *Hypothesis_driven_XGR* was perceived as less predictable compared to *Soft_XGR* ($p = 0.05$) and *Hard_XGR* ($p = 0.03$). Since the Hypothesis_driven model focuses on presenting evidence rather than making direct decisions, the decision-making process may seem more open-ended and variable, affecting the perception of predictability. Although participants' behavioral trust, as reflected in their decision accuracy, is significantly better for our model, further interaction may be necessary to promote perceived trust.

5.5 Explanation Satisfaction

We next present the results of the self-reported satisfaction scale. Considering four metrics—*understand*, *satisfying*, *sufficient_detail*, and *complete*—we obtained p-values of 0.19, 0.11, 0.07, and 0.10 for each metric, respectively. These values indicate marginally significant differences associated with the type of explanation model across most metrics, except for *understand* (see Figure 5b for Likert scale results). Post-hoc pairwise comparisons showed that the explanations provided by the *Hard_XGR* model were perceived as more satisfying, sufficient, and complete compared to those from *Hypothesis_driven_XGR*, with p-values of (0.05, 0.03, 0.06), respectively. *Hard_XGR*, by offering causal relationships, seems to provide a more compelling narrative, which increases satisfaction as users perceive it to align better with their natural desire for understanding why a decision was made, rather than just seeing supporting or opposing evidence. Future improvements could focus on how evidence is presented, potentially incorporating causal links or interactive elements, to ensure explanations are perceived as more satisfying and complete while still promoting user engagement with the decision-making process.

6 CONCLUSION

In this paper, we introduced *Hypothesis-driven* XGR, a model that generates evidence to support and refute goal hypotheses. Our proposed model demonstrates notable improvements in participants' decision-making accuracy, efficiency, and reliance when navigating complex, high-stakes scenarios. This highlights the model's effectiveness in managing challenging tasks and underscores its potential to improve decision-making in complex, real-world settings.

One limitation of our work is that the experiments were conducted with a single type of stakeholder and within a specific domain. Future research should include diverse stakeholder groups and explore other domains, such as healthcare monitoring, traffic management systems, and security surveillance. We also plan to extend our work to an interactive setting, enabling decision-makers to engage with the presented evidence through questions, clarifications, and exploration of alternatives.

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.

REFERENCES

[1] Abeer Alshehri, Tim Miller, and Mor Vered. 2023. Explainable goal recognition: a framework based on weight of evidence. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 33. 7–16.

[2] Abeer Alshehri, Tim Miller, Mor Vered, and Hajar Alamri. 2021. Human centered explanation for goal recognition system.. In *IJCAI-PRICAI Workshop On Explainable Artificial Intelligence (XAI) 2020*. Association for the Advancement of Artificial Intelligence (AAAI).

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[4] Cillian Brewitt, Balint Gyevnar, Samuel Garcin, and Stefano V Albrecht. 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)*. IEEE, 1023–1030.

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

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[7] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[8] Zack Butler, Robert Fitch, Daniela Rus, and Yuhang Wang. 2002. Distributed goal recognition algorithms for modular robots. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, Vol. 1. IEEE, 110–116.

[9] Federico Cabitza, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi. 2024. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine* 150 (2024), 102819.

[10] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[11] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[13] Lee Cordner, Lee Cordner, and Roughley. 2017. *Maritime security risks, Vulnerabilities and cooperation*. Springer.

[14] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

[15] Grady Fitzpatrick, Nir Lipovetzky, Michael Papasimeon, Miquel Ramirez, and Mor Vered. 2021. Behaviour recognition with kinodynamic planning over continuous domains. *Frontiers in Artificial Intelligence* 4 (2021), 717003.

[16] John Fox and Sanford Weisberg. 2018. *An R companion to applied regression*. Sage publications.

[17] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 794–806.

[18] Irving John Good. 1985. Weight of evidence: A brief survey. *Bayesian statistics* 2 (1985), 249–270.

[19] Roger L Granada, Ramon Fraga Pereira, Juarez Monteiro, Duncan Dubugras Alcoba Ruiz, Rodrigo Coelho Barros, and Felipe Rech Meneguzzi. 2017. Hybrid activity and plan recognition for video streams. In *Proceedings of the 31st. AAAI Conference: Plan, Activity and Intent Recognition Workshop, 2017, Estados Unidos*.

[20] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, et al. 2023. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial intelligence review* 56, 4 (2023), 3473–3504.

[21] Josiah P Hanna, Arrasy Rahman, Elliot Fosong, Francisco Eiras, Mihai Dobre, John Redford, Subramanian Ramamoorthy, and Stefano V Albrecht. 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)*. IEEE, 7044–7051.

[22] Rakshith MD Hegde and Harish H Kenchannavar. 2019. A survey on predicting resident intentions using contextual modalities in smart home. *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC)* 11, 4 (2019), 44–59.

[23] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *ArXiv* abs/1812.04608 (2018). https://api.semanticscholar.org/CorpusID:54577009

[24] Yue Hu, Kai Xu, Budhitama Subagdja, Ah-Hwee Tan, and Quanjun Yin. 2021. Interpretable Goal Recognition for Path Planning with ART Networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[25] Gal Kaminka, Mor Vered, and Noa Agmon. 2018. Plan recognition in continuous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[26] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[27] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1369–1385.

[28] Thao Le, Tim Miller, Ronal Singh, and Liz Sonenberg. 2024. Towards the new XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence. *arXiv preprint arXiv:2402.01292* (2024).

[29] Christian Leibig, Moritz Brehmer, Stefan Bunk, Danalyn Byng, Katja Pinker, and Lale Umutlu. 2022. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *The Lancet Digital Health* 4, 7 (2022), e507–e519.

[30] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[31] Peta Masters and Mor Vered. 2021. What's the context? implicit and explicit assumptions in model-based goal recognition. In *International Joint Conference on Artificial Intelligence 2021*. Association for the Advancement of Artificial Intelligence (AAAI), 4516–4523.

[32] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 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*, Vol. 9. 35–47.

[33] Felipe Meneguzzi and Ramon Fraga Pereira. 2021. A Survey on Goal Recognition as Planning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4524–4532. https://doi.org/10.24963/ijcai.2021/616 Survey Track.

[34] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[35] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.

[36] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through friction: an approach for calibrating trust in explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, 1–5.

[37] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 340–350.

[38] Ramon Fraga Pereira, Mor Vered, Felipe Meneguzzi, and Miquel Ramírez. 2019. Online probabilistic goal recognition over nominal models. In *International Joint Conference on Artificial Intelligence 2019*. Association for the Advancement of Artificial Intelligence (AAAI), 5547–5553.

[39] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[40] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[41] Miguel Ramírez and Hector Geffner. 2010. Probabilistic plan recognition using off-the-shelf classical planners. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 24. 1121–1126.

[42] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. 2023. Towards effective human-ai decision-making: The role of human learning in appropriate reliance on ai advice. *arXiv preprint arXiv:2310.02108* (2023).

[43] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[44] Maayan Shvo and Sheila A McIlraith. 2020. Active goal recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9957–9966.
[45] Gita Sukthankar, Christopher Geib, Hung Bui, David Pynadath, and Robert P Goldman. 2014. *Plan, activity, and intent recognition: Theory and practice*. Newnes.
[46] Ivan I Vankov. 2023. The hazards of dealing with response time outliers. *Frontiers in Psychology* 14 (2023), 1220281.
[47] Mor Vered, Piers Howe, Tim Miller, Liz Sonenberg, and Eduardo Velloso. 2020. Demand-driven transparency for monitoring intelligent agents. *IEEE Transactions on Human-Machine Systems* 50, 3 (2020), 264–275.
[48] Mor Vered and Gal A Kaminka. 2017. Heuristic online goal recognition in continuous domains. *arXiv preprint arXiv:1709.09839* (2017).
[49] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. 2023. The effects of explanations on automation bias. *Artificial Intelligence* 322 (2023), 103952.
[50] Grace Wahba. 2002. Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences* 99, 26 (2002), 16524–16530.
[51] Kai Xu, Kaiming Xiao, Quanjun Yin, Yabing Zha, and Cheng Zhu. 2017. Bridging the Gap between Observation and Decision Making: Goal Recognition and Flexible Resource Allocation in Dynamic Network Interdiction.. In *IJCAI*. 4477–4483.
[52] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, et al. 2023. Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems* 3, 3 (2023), 161–188.
[53] Zelun Tony Zhang, Sebastian S Feger, Lucas Dullenkopf, Rulu Liao, Lukas Süsslin, Yuanting Liu, and Andreas Butz. 2024. Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process. *arXiv [cs.HC]* (June 2024). arXiv:2406.08959 [cs.HC] http://arxiv.org/abs/2406.08959