

A Causal Approach to Predicting and Improving Human Perceptions of Social Navigation Robots

Downloaded from: https://research.chalmers.se, 2026-11-18 01:33 UTC

Citation for the original published paper (version of record):
Diehl, M., Tsoi, N., Chavez, G. et al (2026). A Causal Approach to Predicting and Improving Human Perceptions of Social Navigation Robots. IEEE Robotics and Automation Letters, In Press.
http://dx.doi.org/10.1109/LRA.2026.3747547

N.B. When citing this work, cite the original published paper.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, or reuse of any copyrighted component of this work in other works.

(article starts on next page)

A Causal Approach to Predicting and Improving Human Perceptions of Social Navigation Robots

Maximilian Diehl, Nathan Tsoi, Gustavo Chavez, Karinne Ramirez-Amaro, *Member, IEEE*, and Marynel Vázquez, *Member, IEEE*

***Abstract*—As mobile robots are increasingly deployed in human environments, enabling them to predict how people perceive them is critical for socially adaptable navigation. Predicting perceptions is challenging for two main reasons: (1) HRI prediction models must learn from limited data, and (2) the obtained models must be interpretable to enable safe and effective interactions. Interpretability is particularly important when a robot is perceived as incompetent (e.g., when the robot suddenly stops or rotates away from the goal), as it allows the robot to explain its reasoning and identify controllable factors to improve performance, requiring causal rather than associative reasoning. To address these challenges, we propose a Causal Bayesian Network designed to predict how people perceive a mobile robot's competence and how they interpret its intent during navigation. Additionally, we introduce a novel method to improve perceived robot competence employing a combinatorial search, guided by the proposed causal model, to identify better navigation behaviors. Our method enhances interpretability and generates counterfactual robot motions while achieving comparable or superior predictive performance to state-of-the-art methods, reaching an F1-score of 0.78 and 0.75 for competence and intention on a binary scale. To further assess our method's ability to improve the perceived robot competence, we conducted an online evaluation in which users rated robot behaviors on a 5-point Likert scale. Our method statistically significantly increased the perceived competence of low-competent robot behavior by 83%.**

***Index Terms*—Social HRI, human-aware motion planning, acceptability and trust.**

I. Introduction

ROBOTS are becoming more capable, bringing us closer to a future where they will be commonplace in people's lives; however, robots still lack Theory of Mind capabilities [1] that facilitate user adoption of the technology. Specifically, it is crucial for robots to model how their performance is perceived by their users during real-world deployments. For example, consider a mobile robot that guides museum visitors to exhibits [2]. When the robot is perceived as easy to understand, competent, and predictable, its likelihood of adoption increases [3].

Traditionally, human perceptions of robots are measured by surveying users (e.g., with the Robot Social Attribute Scale [4] or the Perceived Social Intelligence Scale [5]), but conducting surveys requires interrupting the flow of interactions and is expensive and time-consuming [6]. As an example, the SEAN Together dataset [3] with 2,964 interaction samples between a mobile robot and a human in a Virtual Reality (VR) environment took months to collect. This number of samples is considered large in Human-Robot Interaction (HRI), yet, small for machine learning standards. These challenges motivated recent work on building more scalable, data-driven models that predict human perceptions of robots [3], [7], [8].

Unfortunately, existing supervised learning approaches used to predict human perceptions of robots [3], [7], [8] only model associations among observed features of interactions, making them prone to spurious correlations in the data [9]. For example, an associative model might suggest that high competence ratings are correlated with a human standing close to the robot, but moving the person closer may not actually improve their perceptions of the robot.

We therefore propose a causal model to predict how humans perceive robots in social navigation scenarios. Causal models are capable of incorporating domain knowledge in their structure, which can reduce the amount of data needed for learning. Moreover, causal models are interpretable and can be used to explain the robot's understanding of the interaction. This allows identifying and reasoning about the causes of a potential failure, as demonstrated in recent studies [10], [11], [12]. Our contributions are three-fold:

1. *An interpretable Causal Bayesian Network (CBN) for predicting human perceptions of robot navigation:* Our CBN maps a sequence of observations, such as the robot's trajectory, to a person's reported impression of the robot's performance along two dimensions: (i) perceived competence and (ii) understandability of the robot's intention. To our knowledge, this is the first time causal modeling has been used to predict human perceptions of robots.
2. *A novel method for generating counterfactual robot trajectories:* When the CBN predicts low perceived competence (Fig. 1(a)), our method generates a counterfactual trajectory via combinatorial search to identify the closest trajectory expected to lead to high perceived competence

Received 15 March 2026; accepted 14 September 2026. Date of publication 26 October 2026; date of current version 5 November 2026. This article was recommended for publication by Associate Editor P. T. Singamaneni and Editor S. Chernova upon evaluation of the reviewers' comments. This work was supported in part by Swedish Research Council under Grant 2025-06377, in part by the National Science Foundation under Grant IIS-2143109 and Grant IIS-1924802, and in part by Barbro Osher Endowment stipend under Grant 90401181. *(Maximilian Diehl and Nathan Tsoi contributed equally to this work.) (Corresponding author: Karinne Ramirez.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Yale Institutional Review Board under Application No. IRB # 2000028891.

Maximilian Diehl and Karinne Ramirez-Amaro are with the Chalmers Univ. of Technology, 41296 Gothenburg, Sweden (e-mail: diehlm@chalmers.se; karinne@chalmers.se).

Nathan Tsoi is with the The University of Texas at Austin, Austin, TX 78712 USA (e-mail: nathan.tsoi@utexas.edu).

Gustavo Chavez and Marynel Vázquez are with the Yale University, New Haven, CT 06511 USA (e-mail: gustavo.chavez@yale.edu; marynel.vazquez@yale.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2026.3747547, provided by the authors.

Digital Object Identifier 10.1109/LRA.2026.3747547

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

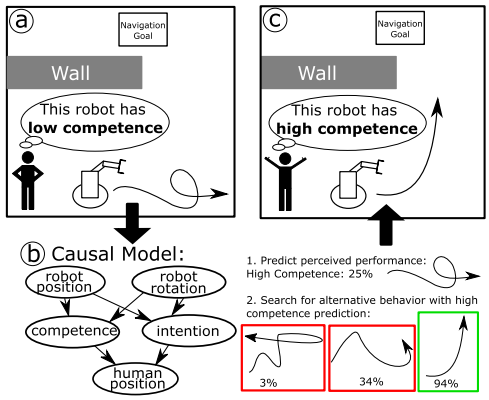


Fig. 1. Example of low-competence robot behavior (a). Our causal model (b) predicts perceived competence by analyzing environmental cues such as the robot's trajectory. When low competence is predicted, the model identifies the minimal environment change (e.g., robot behavior) that is expected to lead to higher competence (c).

(Fig. 1(b)). The robot can then execute this trajectory to preemptively avoid low-competence behavior (Fig. 1(c)). To accommodate the time-series nature of the data (e.g. SEAN Together dataset), we extend prior work on causal models [10], [11] by integrating trajectory clustering. We also refine the search process to exclude infrequently observed, situation-specific behaviors via thresholding.

3. *Validation of our method's ability to generate alternative robot behaviors that improve perceived robot performance through an online user study:* The study shows that the generated counterfactual behaviors can significantly enhance the perceived competence of a mobile robot.

II. Related Work

A. Causality in HRI

Prior work on causal discovery in robotics has primarily focused on manipulation tasks, including tool-affordance learning [13] and cube stacking using Causal Bayesian Networks with physics simulations for next-best-action reasoning [14]. In contrast, causal modeling in HRI remains underexplored. Existing work has applied causal time series analysis to human–robot motion [15] or improving a robot's causal understanding by allowing it to ask humans causal questions [16]. In contrast, our work predicts human perceptions of robot navigation and further uses causal models to generate counterfactual robot behaviors that can help prevent low perceived competence.

B. Failure Prediction and Prevention in Robotics

Our work relates to failure prediction and prevention [17], as low perceived competence and intention in robot behaviors can be seen as interaction failures that we aim to prevent.

1) Predicting Perceived Agent Performance: Recent work utilized machine learning models such as Support Vector Machines, Random Forests, K-Nearest Neighbors, and Multi-Layer Perceptrons to predict a robot's perceived performance [3], [8], user preferences for an agent's helping behavior based on nonverbal cues [7], or models that leveraged social responses to robot errors for timely error detection in human-robot interactions [18]. However, such models have limited interpretability. Moreover, these methods model correlations rather than causal relationships, which can lead to incorrect predictions when used to intervene in the environment to improve human perceptions of robots [9]. We address this limitation by proposing an interpretable Causal Bayesian Network (CBN) for predicting how humans perceive a robot's competence and intention.

2) Failures in Robot Manipulation: FINO-Net [19] proposed a multimodal sensor fusion network to detect and classify failures during manipulation tasks using raw sensory data. REFLECT [20] used multi-sensory data and a large language model (LLM) to generate failure explanations and create correction plans. However, these methods did not consider human perception of robot performance, which are Theory of Mind constructs [1]. Closely related to our contrastive trajectory-generation method, Brandão et al. [21] explained motion-planning failures using contrastive, initialization-based, and trajectory-contrastive techniques. However, their approach lacks a causal model and cannot address perceptions of robot performance in HRI, as it focuses solely on reachability. In contrast, our approach explicitly models human perceptions of robot performance by predicting perceived competence and intention in social navigation scenarios.

3) Failures in Navigation: Robot motion planning is classically defined as finding a collision-free path [22]. Recent approaches extend this definition with predicted human motion to ensure safe navigation [23], [24]. As robots increasingly operate in human environments, it becomes important to account for how humans perceive them [6]. We address this by generating robot trajectories based on perceived competence and intention. Related work on legible motion also incorporates human perception, but relies on hand-crafted cost functions [25] or uninterpretable learning models [26]. In contrast, our CBN learns trajectory classifications directly from data, remaining interpretable, and predicting perceptions along two key dimensions for navigation (competence and intention).

III. Navigation Task

We study the Robot-Following task from the SEAN Together dataset [3], in which a robot guides a VR-controlled human through a warehouse environment with obstacles and other moving agents. The dataset includes observations of nearby autonomously controlled **agents** (within 7.2-m, typically considered the robot's public space [27]) and the **human follower**, whose states are represented by relative positions and orientations (x, y, θ) with respect to the robot. In addition, it contains the **goal position**, specified by relative (x, y) coordinates, and a local **occupancy map** (a 7.2-m $\times$ 7.2-m crop centered on the robot) encoding nearby static obstacles via a ResNet-18embedding. Additionally, the human-follower was asked to rate *how* **competent***the robot was at navigating* and *how clear the robot's* **intentions***were during navigation* on a 5-point Likert scale.[1] To collect these ratings, the navigation was paused at random intervals, and a rating screen was displayed within the VR environment. In our work, we want to modify the robot's behavior when the person perceives the robot as performing

[1]The dataset also contains ratings for how *surprising the robot's navigation behavior was*, which is not considered in our work, as surprise can be interpreted positively or negatively, depending on the context.

poorly (e.g., low competence). We therefore transformed the 5-point format to a binary rating (1–3 = low (0), 4–5 = high (1)). SEAN Together [3] has 2,964 samples from 60 people, $\mathcal{D} = (o_{1:T}, y_i)$. Each sample has 8-second time-series data over all previously described observation variables ($o_{1:T}$) and a single performance rating gathered at the end of that sequence (y_i). We propose a CBN to infer y_i given $o_{1:T}$.

IV. Method

A. Causal Bayesian Networks

Formally, CBNs are defined as *Directed Acyclic Graphs* (DAG) $\mathcal{G} = (\boldsymbol{X}, A)$, where the nodes $\boldsymbol{X} = \{X_1, X_2, \ldots, X_N\}$ are a set of N random variables $X_i \subseteq \boldsymbol{X}$, and A is the set of arcs [28] that describe the causal connections between the variables. We model the Robot-Following task as a CBN, where $\boldsymbol{X}$ represents the task variables as specified in Section IV-B. Based on the dependency structure of the DAG and the *Markov property*, the *joint probability distribution* of a causal BN can be factorized into a set of *local probability distributions*, where each X_i only depends on its direct parents Pa_{X_i}:

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{n} P\left(X_i | \text{Pa}_{X_i}\right) \tag{1}$$

Traditionally, CBNs are obtained in two steps: *Structure Learning* and *Parameter Estimation*.

1) Structure Learning: With sufficient data, the causal structure of a CBN $\mathcal{G} = (\boldsymbol{X}, A)$ can be learned using algorithms such as Grow–Shrink, constraint-based methods that utilize statistical tests to identify conditional independence relations [28]. However, learning plausible causal relationships is challenging [29], particularly with limited data, as in our HRI setting. In such cases, the causal graph is typically specified by a domain expert. The advantage of the separation in structure learning and parameter learning is that the expert only needs to provide high-level causal relationships rather than probability distributions.

2) Parameter Estimation: Once the structure $\mathcal{G}$ is defined, the local conditional distributions $P(X_i|\text{Pa}_{X_i})$ (1) are typically estimated by Maximum Likelihood Estimation (MLE) or Bayesian estimation [30]. In our case, all variables X_i are multinomial, thus each local distribution can be represented as a table of parameters $\boldsymbol{\theta}$. Each entry $\theta_{x|\boldsymbol{u}} \in \boldsymbol{\theta}$ gives the probability $\theta_{x|\boldsymbol{u}} = P(X_i = x|\text{Pa}_{X_i} = \boldsymbol{u})$ of X_i taking value x for a particular parent configuration $\boldsymbol{u}$. The size of this table grows with the number of parents and their number of discretization intervals.

3) Preventing Robot Failures: Our method for preventing low perceived competence builds on prior work that uses CBNs to prevent robot task-execution failures [10], [11], [12]. The core idea is to use the CBN to predict the probability of a desirable outcome, e.g., high perceived competence, based on the current values of the variables that causally impact the competence. If this predicted probability falls below a threshold ϵ, we search for alternative parent-variable values that would yield high perceived competence. However, the existing approaches cannot be directly applied to the SEAN Together dataset for two main reasons. First, the dataset contains time-series variables, such as the robot's trajectory and orientation, whereas prior methods assume all variables are single-valued and discretized. Second, unlike the previous work, which relied on large-scale simulated datasets to learn both the causal structure and the associated conditional distributions, the SEAN Together dataset contains only a limited number of real-world human-robot interactions. This makes it substantially more difficult to infer a reliable causal structure and to estimate the conditional distributions directly from data. In the following subsections, we present our proposed method and discuss how it addresses these challenges.

Algorithm 1: GetVariableIntervals function.

```
1: function: GetVariableIntervals(D, Λ)
2:   X_int ← []
3:   for all i ∈ |X| do
4:     if D_i is not categorical then
5:       if D_i is time-series then
6:         X_int_i ← Add(X_int_i, Cluster(D_i, Λ_i))
7:       else
8:         X_int_i ← Add(X_int_i, Discretize(D_i, Λ_i))
9:     else
10:      X_int_i ← Add(X_int_i, Val(D_i))
11:  return X_int
```

B. Proposed CBN for the Robot-Following Task

1) Data Preprocessing: Our modeling process had three objectives: (1) limit the number of parent variables per node to keep parameter estimation feasible, (2) ensure each node has a clear semantic meaning for interpretability, and (3) make nodes, particularly those influencing competence and intention, actionable by the robot. To achieve this, we manually applied three key modifications to the existing variables. First, all distance variables originally measured in (x, y) coordinates were combined into a single L^2 norm, applied to both robot-goal and human-robot distances. Second, we converted all distances, originally measured as absolute values, into relative changes with respect to the first measurement in each 8-second time series. In this formulation, each distance time-series trajectory begins at 0 (representing the current location) and indicates how the distance changes over the 8-second interval. This ensures all trajectories are physically executable. For example, if our method were to recommend an alternative robot trajectory with a different initial distance to the goal, executing such a trajectory would be infeasible without "teleporting" the robot to a new initial state. Third, variables related to autonomous pedestrians and map information were removed. While this variable selection remains manual and task-dependent, it reduces model complexity and improves interpretability. The final set of variables is shown in Table I.

2) Generalized Discretization Through Time-Series Clustering: A major contribution of our work is increasing the flexibility of existing models to incorporate both time-varying and static variables. To achieve this, we propose a generalized discretization process (Algorithm 1) that can discretize all variables defined in Table I.

The inputs to Algorithm 1 are the training dataset $\mathcal{D}$ (consisting of K IID samples $\boldsymbol{\xi}$, each a fully observed instance of all network variables $\boldsymbol{X}$), and Λ, which specifies the number of discretization intervals per variable. The output is a list of discretization intervals X_{int} for each variable.

For time-series variables $\mathcal{D}_i$ ($\mathcal{D}_i$ denotes all observed values of a specific variable X_i) we apply K-means clustering (Algorithm 1, Line 6), treating each sequence as a vector of data points and grouping similar series by Euclidean distance into Λ_i

TABLE I
CBN VARIABLES $\boldsymbol{X}$. $\theta_t^{\text{ROBOT_GOAL}}$ DENOTES THE ANGLE BETWEEN ROBOT AND GOAL AND DIST_t DENOTES THE EUCLIDEAN DISTANCE AT TIME t.

Variable Name	Formula	Interpretation & Type
robot_rotation_change	$\{\lvert\theta_0^{\text{robot_goal}}\rvert - \lvert\theta_0^{\text{robot_goal}}\rvert, \ldots, \lvert\theta_8^{\text{robot_goal}}\rvert - \lvert\theta_0^{\text{robot_goal}}\}\rvert$	Is the robot rotating towards/away from the goal? (time-series)
total_robot_rotation	$\sum_{t=0}^{8} \lvert\theta_t^{\text{robot_goal}}\rvert$	Total robot rotation (over 8 second observation) (continuous)
initial_robot_rotation	$\theta_0^{\text{robot_goal}}$	Initial robot-goal angle (continuous)
robot_pos_change	$\{\text{dist}_0^{\text{robot_goal}} - \text{dist}_0^{\text{robot_goal}}, \ldots, \text{dist}_8^{\text{robot_goal}} - \text{dist}_0^{\text{robot_goal}}\}$	Is the robot moving towards/away from the goal? (time-series)
competence	$\{0,1\}_{t=8}$	Perceived competence at the end of an observation (categorical)
intention	$\{0,1\}_{t=8}$	Perceived intention at the end of an observation (categorical)
human_pos_change	$\{\text{dist}_0^{\text{human_robot}} - \text{dist}_0^{\text{human_robot}}, \ldots, \text{dist}_8^{\text{human_robot}} - \text{dist}_0^{\text{human_robot}}\}$	Is human moving towards/away from the robot? (time-series)

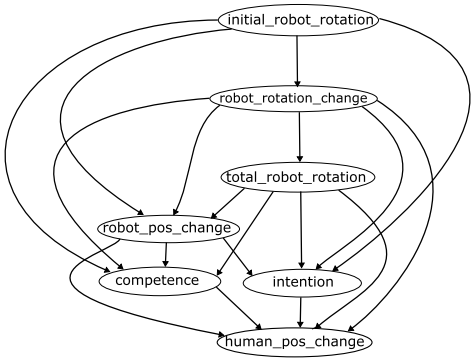


Fig. 2. Proposed CBN graph for the Robot-Following task.

clusters $\{I_1, \ldots, I_{\Lambda_i}\}$. The cluster centroids $\{c_1, \ldots, c_{\Lambda_i}\}$ then define the discretization intervals, representing average patterns. For continuous single-valued variables, we perform quantile discretization [28] (Algorithm 1, Line 8), dividing the data into Λ_i intervals with equal numbers of data points. Categorical variables are discretized by directly assigning the unique values as intervals (Algorithm 1 Line 10). For inference at runtime, each time-series variable (trajectory over the previous 8 seconds) is assigned to the nearest cluster centroid using Euclidean distance.

3) Proposed CBN: Our proposed CBN structure $\mathcal{G}$ for the Robot-Following task is visualized in Fig. 2. Its variables $\boldsymbol{X}$ are defined as in Table I and discretized via Algorithm 1.Our model includes three different aspects of the robot's rotation: The key variable describing the rotational trajectory is `robot_rotation_change`. However, this alone is insufficient for reliably estimating competence or intention, as it measures rotation relative to the start of an 8-second interval and does not capture the initial misalignment toward the goal as measured through `initial_robot_rotation`. For instance, a robot already facing the goal should maintain orientation, whereas a misaligned robot should rotate toward it. We also observed that our clustering struggled to distinguish trajectories that maintain orientation from those where the robot rotates around its own axis, since both have similar net rotation toward the goal. We therefore introduced `total_robot_rotation`, capturing the cumulative rotation over the interval. Another key variable affecting perceived competence and intention is `robot_pos_change`, which represents the robot's movement relative to the goal.

Algorithm 2: Counterfactual Robot Behavior Generation.

Input: discretized failure variable parameterization $x_{\text{current}_{\text{int}}}$, graphical model $\mathcal{G}$, structural equations $P(X_i|\text{Pa}_{X_i})$, discretization intervals of all model variables X_{int}, success threshold ϵ, goal parametrizations X_{goal}
Output: solution variable parameterization $x_{\text{solution}_{\text{int}}}$, solution success probability prediction p_{solution}

1: $P \leftarrow$ GENERATETRANSITIONMATRIX(X_{int})
2: $q \leftarrow [x_{\text{current}_{\text{int}}}]$
3: $v \leftarrow []$
4: **while** $q \neq \emptyset$ **do**
5: $\quad node \leftarrow$ POP(q)
6: $\quad v \leftarrow$ APPEND$(v, node)$
7: $\quad$ **for each** $child$ reachable from $node$ in P **do**
8: $\quad\quad$ **if** $child \notin q, v$ **and** $|\text{Pa}_{\text{child}}| > m$ **then**
9: $\quad\quad\quad p_{\text{solution}} = P(\text{competence} = 1|\text{Pa}_{child})$
10: $\quad\quad\quad$ **if** $p_{\text{solution}} > \epsilon$ **then**
11: $\quad\quad\quad\quad x_{\text{solution}_{\text{int}}} \leftarrow child$
12: $\quad\quad\quad\quad$ RETURN$(p_{\text{solution}}, x_{\text{solution}_{\text{int}}})$
14: $\quad\quad\quad q \leftarrow$ APPEND$(q, child)$

Human movement (`human_pos_change`) is modeled as a consequence of perceived competence, perceived intention, and the robot's motion, where the ratings collected after each 8s interval represent the participant's accumulated perception during the interaction [31].

C. Proposed Method to Address Low Perceived Performance

We propose a new method for failure explanation and prevention (Algorithm 2) that extends the approach of [10], [11] to generate counterfactual robot behaviors. When the current behavior $x_{\text{current}_{\text{int}}}$ is predicted to lead to low competence, Algorithm 2 identifies alternative variable parameterizations that are expected to improve human perception given the CBN.

Line 1 of Algorithm 2 initializes a matrix that defines all valid transitions for each possible parameterization $\boldsymbol{u}$ of the parent variables of the outcome (e.g., competence or intention). The matrix guides the search by allowing only minimal changes, in line with the Occam's Razor principle to find the smallest necessary change. For example, with two parent variables X and Y, each with four intervals ($X = \{x_1, x_2, x_3, x_4\}$ and $Y = \{y_1, y_2, y_3, y_4\}$), a valid transitions from $node = (X = x_1, Y = y_4)$ would be to $child_1 = (X = x_2, Y = y_4)$ or

$child_2 = (X = x_1, Y = y_3)$. These adjust only one variable by one neighboring interval. In contrast, transitions that change multiple variables simultaneously or skip intervals (e.g., x_1 directly to x_3) are disallowed.

In the context of our robot-follower task, a variable parameterization refers to a specific combination of the initial robot rotation interval, total robot rotation interval, rotation cluster, and position cluster ($\text{Pa}_{\text{competence}}$). In our implementation, we prioritize changes to the `robot_pos_change`, `robot_rotation_change`, and `total_robot_rotation` variables before modifying `initial_robot_rotation`. This means our algorithm first searches for possible interventions on robot behavior within the subset of variables that maintain the same initial rotational difference. Only if no successful alternatives are found within this subset does the algorithm consider other initial rotations. To identify the optimal counterfactual behavior, we adapt a Breadth-First Search (BFS) procedure that explores variable parameterizations in search of one that fulfills the target competence or intention criteria. The search terminates upon finding a parameterization $child$ that satisfies the success condition $P(\text{competence} = 1|\text{Pa}_{child}) > \epsilon$, where ϵ represents a threshold indicating a sufficiently high likelihood of achieving the desired competence. This is done in Lines 4-13 of Algorithm 2. Critically, to more reliably deal with small datasets, Line 8 introduces a modification where a parameterization is only considered if it has been observed a sufficient number of times in the training dataset $\mathcal{D}$. Specifically, the search condition is entered only if the count of observations for the parent configuration Pa_{child}, denoted $|\text{Pa}_{child}|$ exceeds a threshold m. This requirement ensures that the algorithm relies only on parameterizations with sufficient support in the data, thereby filtering out infrequent or anomalous behaviors that may not generalize well. Based on a sensitivity analysis, we selected $m = 5$, the smallest threshold yielding a stable number of retained high-competence trajectories over several consecutive values ($m = 5 - -8$), maximizing retained trajectories while filtering low-support behaviors. We set $\epsilon = 0.9$ to ensure a high success probability among the counterfactual solutions.

The final output of Algorithm 2 is a new variable parameterization, $x_{\text{solution}_{\text{int}}}$, for which the model predicts a high rating (e.g., positive competence). This solution provides alternative robot position and trajectory behaviors in the form of cluster centroids and intervals. By implementing these behaviors on the robot, it is expected to be perceived more positively.

V. Clustering and Prediction Performance

In this section, we first describe the clusters obtained from the time-series data provided by SEAN Together [3] and then compare the prediction performance of our Causal Model with that of a Random Forest classifier. The Random Forest was previously identified as the best-performing baseline model among several alternatives for this data [3], including a Graph Neural Network and a Multi-Layer Perceptron.

We evaluate all models using F1-Score, Accuracy, Precision, and Recall. These metrics were computed using Leave-One-Out Cross-Validation (LOOCV). In each run, we held out all samples from one participant for testing and trained the model on the remaining participants. The final performance values are the macro-averaged results across all participants. For the Causal Model, we performed an additional hyperparameter search over the number of intervals Λ using nested LOOCV. With 60 participants, we exhaustively iterated over all splits of 58 for training, 1 for validation, and 1 for testing. The best hyperparameters were selected based on the macro-averaged F1-Score. The final reported results were obtained by running LOOCV once more using these selected hyperparameters.

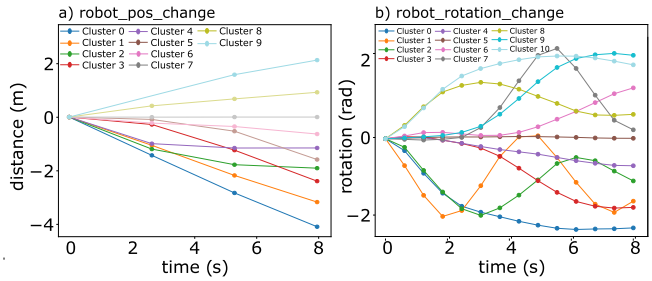


Fig. 3. Cluster means: robot_pos_change and robot_rotation_change.

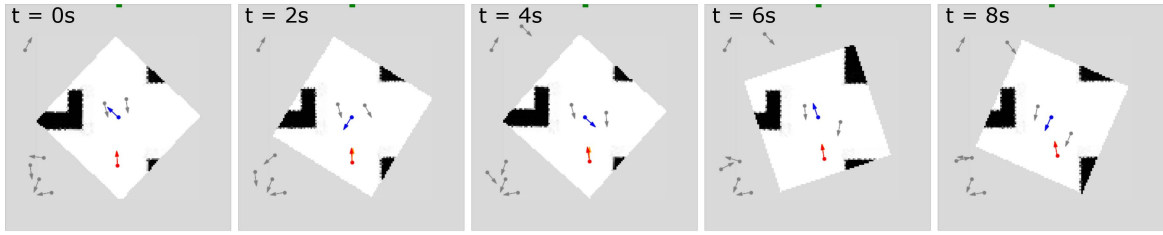
TABLE II
LOOCV Evaluation ($\mu \pm \sigma$) on Binary F1-Score, Accuracy, Precision, and Recall. Our Causal Model Is Compared Against the Best-Performing RF Baseline From Related Work (With Feature Set Nav. + Facial). Bold Indicates the Highest Performance.

Method	F1	Accuracy	Precision	Recall
Competence				
Causal (ours)	$\mathbf{0.777 \pm 0.09}$	$\mathbf{0.835 \pm 0.08}$	0.811 ± 0.12	$\mathbf{0.768 \pm 0.14}$
RF	0.73 ± 0.12	0.814 ± 0.09	$\mathbf{0.816 \pm 0.13}$	0.686 ± 0.16
Intention				
Causal (ours)	$\mathbf{0.751 \pm 0.1}$	$\mathbf{0.788 \pm 0.10}$	$\mathbf{0.823 \pm 0.11}$	$\mathbf{0.713 \pm 0.15}$
RF	0.707 ± 0.12	0.759 ± 0.12	0.817 ± 0.11	0.654 ± 0.18

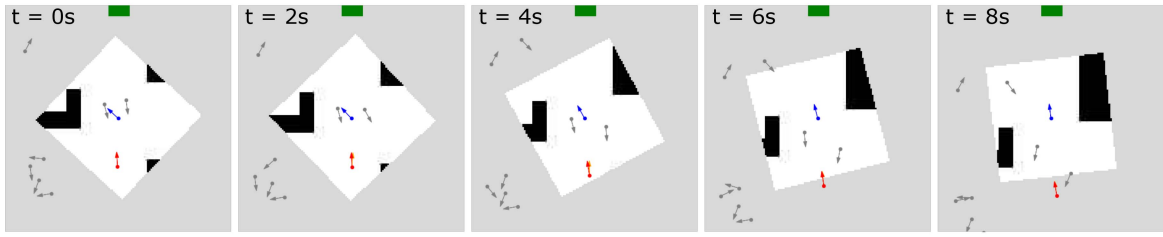
We obtained four intervals for `initial_robot_rot`, `total_robot_rot`, and ten and eleven clusters for `robot_pos_change` and `robot_rotation_change`, respectively (Fig. 3) and ordered them according to the sum of their centroid trajectory values. The position and orientation clusters capture intuitive motion patterns, ranging from increasing (Cluster 9, Fig. 3(a)), constant (Cluster 7, Fig. 3(a)), or decreasing distance to the goal (Cluster 0, Fig. 3(a)) to maintaining orientation (Cluster 5, Fig. 3(b)), rotating toward (Cluster 0, Fig. 3(b)) or away from the goal (Cluster 10, Fig. 3(b)).

In Table II, we compare the prediction performance of our causal model with the best-performant Random Forest (RF) classifier (Nav. + Facial features) from [3]. Measured by Accuracy, our model outperformed the baseline by 0.021 for Competence and 0.029 for Intention. Measured by F1-Score our method outperformed the baseline by 0.047 and 0.044 for Competence and Intention, respectively. Statistical significance was assessed using the Nadeau–Bengio corrected paired t-test [32], which accounts for the dependence between cross-validation folds due to overlapping training sets. The causal model significantly outperformed the non-causal baseline in F1 score for both competence ($t(59)=3.29$, $p=0.0017$) and intention ($t(59)=2.58$, $p=0.0125$), as well as in classification accuracy for competence ($t(59)=2.53$, $p=0.0142$) and intention ($t(59)=2.22$, $p=0.0302$).

In conclusion, on average, our method outperformed the prior state-of-the-art, uninterpretable machine learning approach. Our causal model has the additional benefit that it is interpretable and simpler, using only a subset of the full feature set from SEAN Together. Moreover, the model encodes causal information, which is crucial for generating behaviors for improving the robot's perceived performance.



(a) Original robot behavior



(b) Counterfactual robot behavior

Fig. 4. Image series of a navigation-task video for our user study (Section VI). The blue arrow represents the robot's position and orientation, the red arrow depicts the follower, and other agents are shown as grey arrows. The green rectangle at the top of the image is the goal. Black areas indicate static obstacles, white areas are navigable space, and the surrounding grey region lies outside the 7.2 m public space [27] where the 2D map was recorded. The upper series shows the original robot behavior. The robot rotates around its own axis, which was classified as low competence by our model. The lower series depicts the counterfactual behavior generated to address this low-competence trajectory, where the robot rotates towards the goal and then proceeds along a straight line towards it. Images are best viewed in color.

VI. User Study of Counterfactual Behavior Generation

We conducted an online user study to evaluate the capability of our method to generate counterfactual trajectories with greater *perceived robot competence*, using the Prolific crowdsourcing platform. Our study was reviewed and approved by the Yale Institutional Review Board (IRB # 2000028891).

A. Hypotheses

We hypothesized that our method can improve the perceived competence of robot navigation behavior, specifically:

H1: When our causal model predicts the perceived competence *correctly as low*, our approach generates navigation behaviors that are perceived as more competent than the original robot behavior.

H2: When our causal model predicts the perceived competence *erroneously low*, our approach still generates navigation behaviors that are perceived as more competent than the original robot behavior.

To test both hypotheses, we implemented two distinct study phases: One phase included 10 scenarios in which our model correctly predicted low robot competence. The other phase studied 10 scenarios where the model incorrectly classified the robot as low competent. We recruited a different set of participants for each study phase.

B. Study Phases and Conditions

Both study phases included two conditions: "Original", and "Counterfactual." In the Original condition, participants viewed videos directly from the SEAN Together dataset (Fig. 4(a)). In the Counterfactual condition, videos were generated by modifying the robot's behavior to follow the average cluster means of alternative `robot_pos_change` and `robot_rotation_change` clusters as obtained through our proposed method (Fig. 4(b)). The generated counterfactuals represent abstract behavioral specifications rather than executable robot trajectories, allowing the method to identify desired behavioral changes independently of the robot platform or motion planner. For the user study, these specifications are visualized using a simplified playback procedure that abstracts away embodiment and kinematic constraints: Position trajectories (e.g., decreasing the robot's distance towards the goal by $4\mathrm{m}$ as shown by Cluster 0 in Fig. 3(b)) were adjusted along a straight line toward the goal. Rotations were down-scaled if smaller than the cluster maximum but not increased; as a result, in some cases, the robot moved toward the goal without fully aligning its orientation. Position and rotation trajectories were synchronized in time but applied independently. To maintain visual consistency with the original videos, the robot remained centered, while the map, pedestrians, and human follower positions were adjusted. We could not predict how the human follower and pedestrians would respond to the counterfactual robot behavior and, thus, kept their movements unchanged.

C. Experimental Procedure

Participants completed the online study using a standard desktop web browser; mobile devices were not permitted. After providing consent, they completed a brief demographic survey and read study instructions, which explained the SEAN Together dataset [3], the task of rating robot trajectories, the expected duration ($\approx$15 minutes) and payment (USD \$3.00). Participants were familiarized with the top-down video perspective by viewing an example video. Each participant viewed 20 videos in total: 10 showing the original robot behavior and 10 showing counterfactual behaviors. These videos were generated from a random sample of 10 scenarios from a 120-scenario test set not included in training. All participants saw the same 10 scenarios (both original and counterfactual), but the order was randomized to reduce ordering effects. After each video, participants rated

how they believed the human follower perceived the robot's competence on a 5-point Likert scale (from 1 = "very incompetent" to 5 = "very competent"). A video with a visual overview of the study is provided under the supplementary materials of the paper.

For each study phase, we recruited 40 participants (80 total) from North America, aged 18 or older, with normal or corrected-to-normal vision, and fluent in English. Their mean age was 37.21 ± 12.30 years, with 13 female and 27 male participants. On average, participants reported near-neutral familiarity with robots (2.94 ± 1.41 on a 7-point scale, where 1 = "Not at all familiar" and 7 = "Very familiar").

D. Results

We fitted linear mixed-effect models for the competence measure with fixed effects for the condition (Original or Counterfactual). A linear mixed-effect model was used due to the hierarchical nature of the data, i.e., Participant ID was nested within Scenario ID and added as a random effect to the model. This nesting associated the robot's behavior across the two instances of a scenario shown to each participant.

(H1) Competence of the Counterfactuals when the CBN model *correctly* predicted the Original robot behavior as low competent. We found a significant difference in participants' ratings between the original trajectories and our model's proposed counterfactual trajectories when our CBN model made a correct prediction, $F(1, 399) = 338.36, p < 0.0001$. Overall, participants rated the competence of counterfactual trajectories ($3.59 \pm .06$) higher than the original trajectories (1.96 ± 0.06). Therefore, on average, counterfactual trajectories improved the perceived competence of the robot by 83%. These results support H1, demonstrating our method's potential to improve robot trajectories and enhance the perceived competence of a robot.

(H2) Competence of the Counterfactuals when the CBN model *erroneously* predicted the Original robot behavior as low competent. We found a significant difference in participants' ratings between the original trajectories and our model's proposed counterfactual trajectories when our CBN model made an incorrect prediction, $F(1, 399) = 76.32, p < 0.0001$. Overall, participants rated the competence of counterfactual trajectories ($3.42 \pm .07$) higher than the original trajectories (2.69 ± 0.07). Thus, our counterfactual trajectories improved the perceived competence of the robot by 27%. These results support H2, demonstrating our method's potential to improve robot trajectories and enhance the perceived competence of the robot, even when the model makes an incorrect initial prediction regarding the robot's perceived competence.

E. Discussion

Our results suggested that, across both study phases, the counterfactual trajectories were rated significantly more competent than the original robot behaviors. In terms of absolute scores, the competence ratings for the generated counterfactuals were consistent between the two study phases, with an overall average rating of 3.51 over the 5-point competence scale. Although the original robot trajectories in the second study phase (2.69 ± 0.07) were rated higher than in the first phase (1.96 ± 0.06), their average ratings were below the midmark of the competence scale. This suggests that the proposed causal model is capable of identifying additional cases where a correction of the robot's trajectory may be needed for an average user.

To isolate the contribution of the causal graph, we additionally evaluated a non-causal baseline by training an RF predictor on the same causal parent variables of competence as the causal model, with the addition of `human_pos_change`, which is modeled as a downstream consequence rather than a causal parent. We then applied the same BFS procedure based on the non-causal RF to the 20 videos used in the user study. In 11 of the 20 cases, the search recommended modifying the human trajectory. This shows that while the search procedure contributes to finding higher-rated alternatives, the causal model provides the intervention semantics required to distinguish actionable causal interventions from purely associative variables. Additionally, while our current implementation uses MLE, recent results show that using an RF trained on the same causal parent variables improves predictive performance (competence: F1-score 0.890 ± 0.085, accuracy 0.871 ± 0.087; intention: F1-score 0.849 ± 0.129, accuracy 0.845 ± 0.107). Exploring more expressive predictors is an interesting direction for future work.

VII. Limitations & Future Work

To enable our CBN model to effectively learn from small datasets, we excluded environmental factors such as walls and pedestrians from the causal graph. This simplification reduced the model complexity but violated the assumption of no unobserved confounding variables [9]. Our model captures the relationship between robot motion and perceived competence but does not explicitly model the environmental context, such as obstacle avoidance. We observed only minimal negative effects in our study: in one of the 20 evaluated scenarios, the generated counterfactual rotated and moved towards a wall instead of around it. Future work will investigate richer causal graph structures that incorporate environmental context and more complex interaction dynamics, including cyclic dependencies between human and robot behavior.

For the user study, we lacked ground-truth follower behavior for the counterfactual robot trajectories and therefore reused the original human trajectories (e.g., Fig. 4(b)), increasing the gap between the robot and follower and potentially causing participants to perceive the robot as moving too quickly or not coordinating sufficiently with the human. Future work will integrate the high-level counterfactual behaviors into a robot motion planner, enabling high-competence robot behavior while respecting environmental constraints such as obstacles, pedestrian flow, and robot kinematics. This will also enable immersive VR and real-world evaluations with naturally emerging human follower trajectories.

While previous work has learned causal graph structures from sufficiently large datasets (e.g., 20,000 samples) [10], we manually specified the causal graph while learning all CBN probability distributions, preserving interpretability and enabling causal reasoning in HRI. To reduce manual effort when transferring the approach to new domains, future work will explore using Large Language Models to propose candidate causal structures [33], particularly in the data-limited settings common in HRI.

VIII. Conclusion

We proposed a Causal Bayesian Network for online prediction of human perceptions of a navigation robot's competence and intention. The model outperformed the state-of-the-art supervised

correlative model in predicting perceived robot competence and intention while remaining interpretable. We also proposed a new method for generating counterfactual robot behaviors that improved the perceived competence of low-competence robot navigation by 83%.

REFERENCES

[1] Y. Mao, S. Liu, Q. Ni, X. Lin, and L. He, "A review on machine theory of mind," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 6, pp. 7114–7132, Dec. 2024.

[2] Y. Sasaki and J. Nitta, "Long-term demonstration experiment of autonomous mobile robot in a science museum," in *Proc. IEEE Int. Symp. Robot. Intell. Sensors*, 2017, pp. 304–310.

[3] Q. Zhang et al., "Predicting human perceptions of robot performance during navigation tasks," *ACM Trans. Hum.-Robot Interaction*, vol. 14, no. 3, pp. 1–27, 2025.

[4] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2017, pp. 254–262.

[5] K. A. Barchard, L. Lapping-Carr, R. S. Westfall, A. Fink-Armold, S. B. Banisetty, and D. Feil-Seifer, "Measuring the perceived social intelligence of robots," *ACM Trans. Hum.-Robot Interaction*, vol. 9, no. 4, pp. 1–29, 2020.

[6] N. Tsoi, R. Sterneck, X. Zhao, and M. Vázquez, "Influence of simulation and interactivity on human perceptions of a robot during navigation tasks," *ACM Trans. Hum.-Robot Interaction*, vol. 13, no. 4, pp. 1–19, 2024.

[7] K. Candon, J. Chen, Y. Kim, Z. Hsu, N. Tsoi, and M. Vázquez, "Nonverbal human signals can help autonomous agents infer human preferences for their behavior," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2023, pp. 307–316.

[8] Q. Zhang, A. Narcomey, K. Candon, and M. Vázquez, "Self-annotation methods for aligning implicit and explicit human feedback in human-robot interaction," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2023, pp. 398–407.

[9] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, Inc., 2018.

[10] M. Diehl and K. Ramirez-Amaro, "Why did i fail? A causal-based method to find explanations for robot failures," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 8925–8932, Oct. 2022.

[11] M. Diehl and K. Ramirez-Amaro, "A causal-based approach to explain, predict and prevent failures in robotic tasks," *Robot. Auton. Syst.*, vol. 162, 2023, Art. no. 104376.

[12] M. Diehl and K. Ramirez-Amaro, "Generating and transferring priors for causal Bayesian network parameter estimation in robotic tasks," *IEEE Robot. Automat. Lett.*, vol. 9, no. 2, pp. 1011–1018, Feb. 2024.

[13] J. Brawer, M. Qin, and B. Scassellati, "A causal approach to tool affordance learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 8394–8399.

[14] R. Cannizzaro, M. Groom, J. Routley, R. O. Ness, and L. Kunze, "A causal Bayesian network and probabilistic programming based reasoning framework for robot manipulation under uncertainty," 2024, *arXiv:2403.14488*.

[15] L. Castri, S. Mghames, M. Hanheide, and N. Bellotto, "Causal discovery of dynamic models for predicting human spatial interactions," in *Social Robotics*. Switzerland: Springer Nature, 2022.

[16] F. Edström, T. Hellström, and X. De Luna, "Robot causal discovery aided by human interaction," in *Proc. 32nd IEEE Int. Conf. Robot, Hum. Interactive Commun.*, 2023, pp. 1731–1736.

[17] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Front. Psychol.*, vol. 9, 2018, Art. no. 861.

[18] M. Stiber, R. Taylor, and C.-M. Huang, "Modeling human response to robot errors for timely error detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 676–683.

[19] A. Inceoglu, E. E. Aksoy, and S. Sariel, "Multimodal detection and classification of robot manipulation failures," *IEEE Robot. Automat. Lett.*, vol. 9, no. 2, pp. 1396–1403, Feb. 2024.

[20] Z. Liu, A. Bahety, and S. Song, "Reflect: Summarizing robot experiences for failure explanation and correction," in *Proc. 7th Conf. Robot Learn.*, 2023, pp. 3468–3484.

[21] M. Brandão, G. Canal, S. Krivić, and D. Magazzeni, "Towards providing explanations for robot motion planning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 3927–3933.

[22] M. Mohanan and A. Salgoankar, "A survey of robotic motion planning in dynamic environments," *Robot. Auton. Syst.*, vol. 100, pp. 171–185, 2018.

[23] L. Heuer, L. Palmieri, A. Rudenko, A. Mannucci, M. Magnusson, and K. O. Arras, "Proactive model predictive control with multi-modal human motion prediction in cluttered dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 229–236.

[24] S. Schaefer, K. Leung, B. Ivanovic, and M. Pavone, "Leveraging neural network gradients within trajectory optimization for proactive human-robot interactions," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 9673–9679.

[25] J. Amirian, M. Abrini, and M. Chetouani, "Legibot: Generating legible motions for service robots using cost-based local planners," in *Proc. 33rd IEEE Int. Conf. Robot Hum. Interactive Commun.*, 2024, pp. 461–468.

[26] S. Wallkötter, M. Chetouani, and G. Castellano, "SLOT-V: Supervised learning of observer models for legible robot motion planning in manipulation," in *Proc. 31st IEEE Int. Conf. Robot Hum. Interactive Commun.*, 2022, pp. 1421–1428.

[27] W. Jensen, S. Hansen, and H. Knoche, "Knowing you, seeing me: Investigating user preferences in drone-human acknowledgement," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–12.

[28] M. Scutari, "Learning Bayesian networks with the BNlearn R package," *J. Stat. Softw.*, vol. 35, no. 3, pp. 1–22, 2010.

[29] A. Sharma, V. Syrgkanis, C. Zhang, and E. Kiciman, "Dowhy: Addressing challenges in expressing and validating causal assumptions," *Proc. Workshop, Neglected Assumptions, Causal Inference*, 2021.

[30] Z. Ji, Q. Xia, and G. Meng, "A review of parameter learning methods in bayesian network," in *Advanced Intelligent Computing Theories and Applications*. Berlin, Germany: Springer Int. Publishing, 2015.

[31] Z. Liang, A. Mitrevski, E. Dean, and K. Ramirez-Amaro, "Can causal models enhance robot navigation? Online causal adaptation for real-robot navigation," *IEEE 35th Int. Conf. Robot Human Interactive Commun.*, 2026, pp. 1843–1850.

[32] C. Nadeau and Y. Bengio, "Inference for the generalization error," in *Proc. Adv. Neural Inf. Process. Syst.*,, vol. 12, 1999, pp. 307–313.

[33] T. Jiralerspong, X. Chen, Y. More, V. Shah, and Y. Bengio, "Efficient causal graph discovery using large language models," 2024, *arXiv:2402.01207*.