Predicting Human Impressions of Robot Performance During Navigation Tasks

Qiping Zhang†, Nathan Tsoi†, Booyeon Choi†, Jie Tan‡, Hao-Tien Lewis Chiang‡, Marynel Vázquez†
†Yale University ‡Google DeepMind

***Abstract*—Human impressions of robot performance are often measured through surveys. As a more scalable and cost-effective alternative, we investigate the possibility of predicting people's impressions of robot behavior using non-verbal behavioral cues and machine learning techniques. To this end, we first contribute the SEAN TOGETHER Dataset consisting of observations of an interaction between a person and a mobile robot in a Virtual Reality simulation, together with impressions of robot performance provided by users on a 5-point scale. Second, we contribute analyses of how well humans and supervised learning techniques can predict perceived robot performance based on different observation types (like facial expression features, and features that describe the navigation behavior of the robot and pedestrians). Our results suggest that facial expressions alone provide useful information about human impressions of robot performance; but in the navigation scenarios that we considered, reasoning about spatial features in context is critical for the prediction task. Also, supervised learning techniques showed promise because they outperformed humans' predictions of robot performance in most cases. Further, when predicting robot performance as a binary classification task on unseen users' data, the F_1-Score of machine learning models more than doubled in comparison to predicting performance on a 5-point scale. This suggested that the models have good generalization capabilities, although they are better at telling the directionality of robot performance than predicting exact performance ratings. Based on our findings, we provide guidelines for implementing supervised learning models that map implicit human feedback to impressions of robot performance in real-world navigation scenarios.**

I. Introduction

As a scalable alternative to measuring subjective impressions of robot performance through surveys, recent work in Human-Robot Interaction (HRI) has explored using *implicit* human feedback to predict these impressions [2, 18, 57, 72]. These are communicative signals that are unintentionally exhibited by people [36]. They can be reflected in human actions that change the world's physical state [53] or can be nonverbal cues, such as facial expressions [18, 57] and gaze [47, 2], displayed during social interactions. Implicit feedback serves as a burden-free information channel that sometimes persists even when people don't intend to communicate [35].

We expand the existing line of research on predicting impressions of robot performance from nonverbal human behavior to dynamic scenarios involving robot navigation. Prior work has often considered stationary tasks, like physical assembly at a desk [58] or robot photography [72], in laboratory environments. We instead explore the potential of using observations of body motion, gaze, and facial expressions to predict a human's impressions of robot performance while a robot guides them to a destination in a crowded environment.

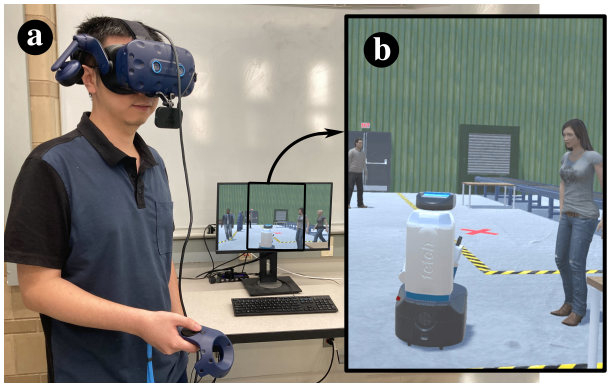

Fig. 1: Data collection. Humans controlled an avatar in the simulation with VR (a) while they were guided by a Fetch robot (b). The screen on the desk shows what the user saw.

These impressions (which we also refer to as human perceptions) correspond to subjective opinions of how well a robot is performing the navigation task. Predicting them in crowded navigation scenarios is more challenging than in stationary settings because human nonverbal behavior can be a result of not only robot behavior, but also other interactants in the environment. Further, because of motion, nonverbal responses to the robot may change as a function of the interaction context. For example, imagine that the person that follows the robot looks downwards. This could reflect paying attention to the robot, or be a result of the person inspecting their nearby physical space, which changes during navigation.

To study implicit feedback during navigation tasks, we performed a systematic data collection using the Social Environment for Autonomous Navigation (SEAN) 2.0 [67] with Virtual Reality (VR) [73].[1] Humans took part in the simulations through an avatar, which was controlled using a VR headset, as in Fig. 1. The headset enabled immersion and allowed us to capture implicit feedback features like gaze. Also, it facilitated querying the human about robot performance as navigation tasks took place. We considered robot performance as a multi-dimensional construct, similar to [72], because humans may care about many aspects of a robot's navigation behavior, as discussed in the social robot navigation literature [22, 44, 21].

Then, we investigated fundamental questions about the value of implicit feedback signals in predicting subjective impressions of robot performance. First, we investigate to what extent humans can predict a person's impression of the

[1]Dataset and code to be released upon acceptance.

performance (along the dimensions of perceived competence, surprise, and intention) based on visualizations of the observations of their interactions, as recorded in our navigation dataset. Second, we investigate how well various supervised learning models do this type of inference in comparison to humans. Finally, we study the generalization capabilities of supervised learning methods to unseen users.

Our analyses bring understanding to the complexity of predicting humans' impressions of robot performance in navigation tasks. Based on our findings, we conclude this paper with a set of suggested guidelines for implementing machine learning algorithms that infer robot performance using implicit feedback in real-world navigation scenarios. We hope that these guidelines facilitate future efforts to make robots more aware of their failures during navigation [63], as well as facilitate aligning robot behavior to human preferences based on implicit feedback [45, 18, 16].

II. Related Work

A. Impressions of Robot Performance

Understanding human impressions of robot performance is important. They can be used to evaluate robot policies [61, 39, 49] and to create better robot behavior [62, 47, 17, 7], increasing the likelihood of robot adoption. In this work, we focus on inferring three robot performance dimensions relevant to navigation [22]: robot competence, surprising behavior, and clear intent. Robot competence is a popular performance metric [14], especially in robot navigation [43, 66, 1]. Surprising behavior violates expectations. It is often considered undesired [3, 21] and may require explanations by the robot [10]. Meanwhile, showing clear intent means that the robot enables an observer to infer the goal of its motion [19]. If humans fail to anticipate the motion of a robot because it acts surprisingly or its intent is unclear, they will likely have trouble coordinating their own behavior with it [54, 20].

B. Implicit Human Feedback

We distinguish between explicit and implicit human feedback about robot performance. Explicit feedback corresponds to purposeful or deliberate information conveyed by humans to robots, e.g., through preferences [8, 60] or survey instruments [4, 43]. Meanwhile, implicit feedback are cues and signals that people exhibit without intending to communicate some specific information about robot performance, yet they can be used to infer such perceptions. Inferring performance from implicit feedback can reduce the chances of excessively querying users for explicit feedback in robot learning scenarios [52, 25], thereby minimizing the risk of feedback fatigue [38]. Learning from implicit feedback is not without challenges, however, as it can be difficult to interpret [18, 57]. For example, this can happen due to inter-person variability in facial expressions [26] or similar signals being produced for different reasons [12].

Our work considers a variety of nonverbal implicit signals, including gaze, body motion, and facial expressions, which have long been studied in social signal processing [69]. While in some cases these signals are treated as explicit feedback (e.g., to interrupt an agent [71]), we consider them implicit feedback because we do not prime humans to react in specific ways to a robot. As such, our work is closer to [18, 70, 45, 56, 12], which used nonverbal signals to identify critical states during robot operation, detect robot errors, and adjust robot behavior.

C. Data Collection in HRI: VR and Other Methodologies

Different kinds of HRI research methods have been used in the literature to gather interaction data, such as in-person user studies (e.g., [23, 65, 43]), observational public data collections (e.g., [42, 34]), crowdsourcing studies (e.g., [11, 64, 32]), etc. See [5] for an introduction to these methods.

We considered different ways of conducting our data collection, but ultimately opted for gathering data with simulated human-robot interactions in VR for several reasons. First, in contrast to real-world data collection, simulation facilitated querying humans about their impressions of robot performance during interactions and resulted in fewer negative consequences for interrupting the navigation task. This is illustrated in Fig. 2. In lab studies, for instance, surveys that gather general impressions of a robot are typically administered at the end of interactions to avoid interrupting the natural flow of events [72], which can cause unintended effects on collaborative tasks and interactants. In VR simulations, however, we can gather feedback in-situ. We can freeze time during human-robot interactions, query a participant about their impressions of robot performance through the VR display, and then resume the simulation as if the interruption had not occurred.

Second, simulations made human-robot interactions safer in contrast to real-world interactions as we wanted to expose participants not only to good robot navigation behavior, but also bad behavior. This was important for inducing a wide range of impressions about robot performance and, thus, capturing varied implicit feedback signals. Prior work has used simulations in HRI for safety reasons as well (e.g., [46, 30]).

Third, in contrast to crowdsourcing, our in-person data collection reduced unrelated participant distractions [9] and minimized potential issues with participant's internet speed [31, 66]. Early in our research process, we considered using interactive surveys [66] for our data collection while capturing implicit feedback signals through the webcams of remote participants (e.g., as in [12]). However, after testing both this setup and VR, we thought that the increased level of immersion afforded by VR was important to gather naturalistic human feedback.

III. Problem Statement & Research Questions

We study if a person's impression of a robot's performance can be predicted using observations of their interaction in dynamic tasks involving navigation. Specifically, we aim to learn a mapping from a sequence of observations to an individual's reported impressions at the end of the sequence (as in Fig. 2b). We consider multiple robot performance dimensions on a 5-point scale, as detailed later in Sec. IV.

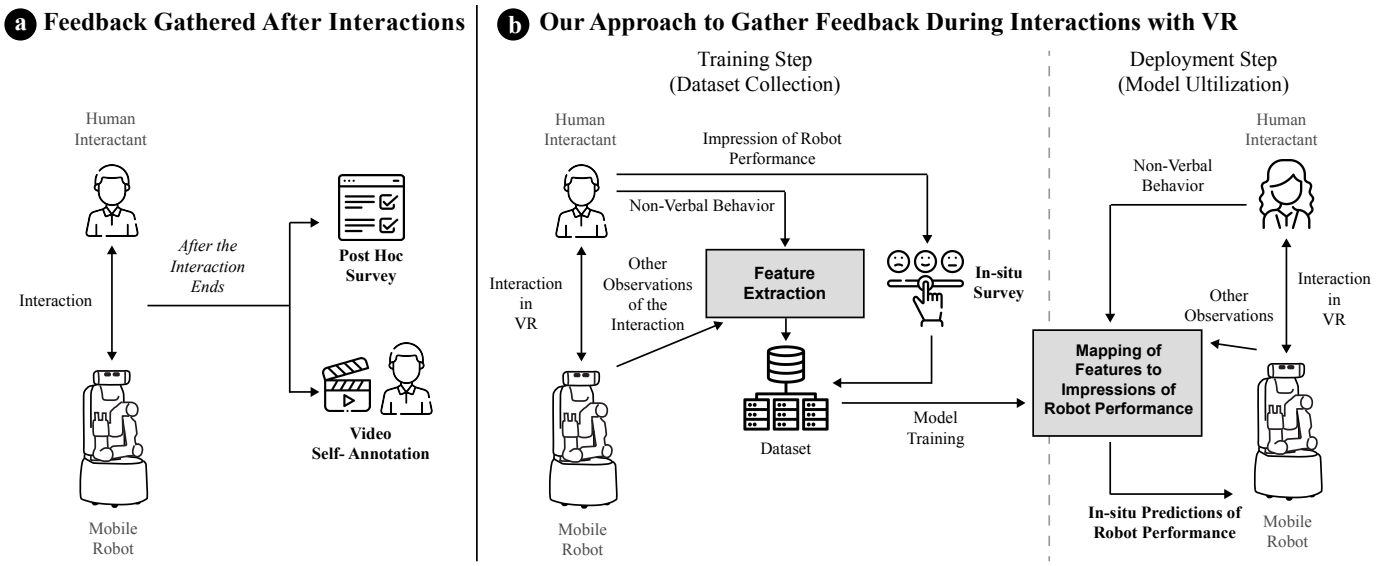

Fig. 2: a) It is typical to gather explicit human feedback about robot performance using surveys after human-robot interactions conclude because interruptions by the experimenters can easily bias the social encounters. Unfortunately, the feedback from surveys tends to be very limited, making it difficult to understand robot performance at a granular level. Alternatively, participants may complete video annotations of their experiences [73], but this can be time consuming and taxing, especially in continuous navigation tasks. b) In this work, we first collect a dataset of human impressions of a robot's performance by prompting participants *during* interactions using VR. Then, we use this explicit feedback to train models that infer human impressions of robot performance based on observations of the interactions, especially including observations of human implicit feedback. The value of such a model is that once it is trained, it can be reused to estimate robot performance during new interactions (Deployment Step), without having to ask humans for explicit feedback anymore.

Consider a dataset of observations and performance labels, $\mathcal{D} = \{(\mathbf{o}^i_{1:T}, y^i)\}$, where $\mathbf{o}_{1:T}$ is an observation sequence of length T, y is a performance rating given by a robot user at the end of the sequence, and i identifies a given data sample. We place emphasis on predicting a person's impression of a robot by considering observations of their implicit feedback. Thus, the observations $\mathbf{o}^i_t$ include features that describe the person's non-verbal behavior, such as gaze and facial expressions. Also, the observations include features that describe the spatial behavior of all the agents in the environment, the navigation task, and the space occupied by static objects. Given this data, we investigate three fundamental research questions:

1) *How well can human observers predict a user's impression of robot performance?* By answering this question, we obtain a human baseline for learning a function $f : \mathcal{O}_{1:T} \rightarrow \mathcal{Y}$, where $\mathcal{O}$ is the observation space and $\mathcal{Y}$ is performance. Also, through this question, we study the impact of two types of observations in the prediction task: observations that describe fine-grained facial expressions for a robot user; and other observations about the user, the robot and their environment. As mentioned earlier, observations of fine-grained expressions have gained popularity in recent work to infer human perceptions of an agent's behavior [18, 12, 72, 58]. Other observations (e.g. body motion and nearby static obstacles) can be more easily computed in real-world navigation tasks, but their usefulness on a robot's ability to infer users' impression of their performance is less understood.

2) *Can machine learning methods predict impressions of robot performance as well as humans?* Ultimately, we are interested in bringing us forward to a future where machine learning models facilitate evaluating robot performance at scale, without having to necessarily ask users all the time for explicit feedback (as in the Deployment Step of Fig. 2b). Thus, we evaluate various machine learning models to approximate the function f, as defined in the prior question.

3) *How well can machine learning models generalize to unseen users?* In future robot deployments, a robot may interact with completely new users. Thus, we conduct a more detailed analysis of the performance of various machine learning models in predicting impressions of robot performance according to users for whom the model had no data at training time.

IV. Data Collection with SEAN and VR

For our data collection, we leveraged the SEAN 2.0 simulator [67]. SEAN 2.0 integrates with the Robot Operating System (ROS) [51] and supports Virtual Reality [73]. Participants used a Vive Pro Eye VR device to control an avatar in a warehouse (as in Fig. 1(a)). The VR headset captured implicit signals from the participants, like eye and lip movements.

During data collection, the participants had to follow a Fetch robot that guided them to a destination that was unknown to them a priori but was marked by a red cross on the ground. Fig. 1(b) shows an example first-person view of the simulation during robot-guided navigation. The Fetch robot was controlled with ROS in SEAN. The environment contained other

algorithmically controlled pedestrians and obstacles typical of warehouses.

The participants provided ratings of robot performance through the simulation's VR interface. The frame rate of the rendering of the virtual environment in the participants' first-person view in VR was over 30 frames per second. Our data collection protocol, described below, was approved by our local Institutional Review Board and refined through pilots.

A. Participants

We recruited 60 participants using flyers and by word of mouth. They were at least 18 years old, fluent in English, and had normal or corrected-to-normal vision. Overall, 19 participants identified as female, 40 as male, and 1 as non-binary or third gender. Most of them were university students, and ages ranged from 18 to 43 years old. Participants were somewhat familiar with robots, as indicated by a mean rating of M = 4.20 (with standard error SE = 0.18) on a 7-point Likert responding format (1 being lowest). Yet, they were somewhat unfamiliar with VR (M = 3.72, SE = 0.20). No participant had prior experience with SEAN or social robot navigation in VR.

B. Data Collection Procedure

Protocol: A data collection session took place as follows. First, the participant provided demographics data. Second, the experimenter introduced the robot, explained the navigation task in which the participant was to follow the robot, and demonstrated how to use the VR device to control their avatar in SEAN and label robot performance. Third, the participant experienced four navigation tasks with the robot, each with a particular starting position and destination. For consistency, the pedestrians were controlled using the same behavior graph controller provided in SEAN 2.0 [67] and the robot used the same navigation logic across the tasks.

In each task, the robot guided the participant to the destination and repeatedly changed its behavior (as further detailed below). Importantly, the interaction was paused before and after each behavior change took place, at which point the participant was asked to evaluate the robot's most recent navigation performance. A typical data collection session was completed in 45 min to 1 hour. Participants were compensated US$15 for their time.

Robot Behaviors: During a navigation task, the robot switched between one of these three types of behavior:

1. Nav-Stack. The robot navigated efficiently to the destination based on the path planned by the ROS Navigation Stack with social costs [40]. This behavior lasted 40 seconds.

2. Spinning. The robot rotated at its current position, which we expected to be perceived as if the robot was confused. This behavior lasted 20 seconds. It was implemented by sending angular velocity commands to the robot's motion controller.

3. Wrong-Way. The robot moved in the wrong direction, away from the task's destination, effectively making a mistake during navigation. This behavior lasted 20 seconds and was implemented using the Navigation Stack as well, but with an incorrect navigation goal.

Unbeknownst to the participants, the robot switched to *Nav-Stack* behavior after *Spinning* or *Wrong-Way* during navigation. It randomly switched to *Spinning* or *Wrong-Way* after finishing *Nav-Stack*. The design was intended to maintain a consistent rate of sub-optimal behavior and avoid user boredom or significant confusion, which can be caused by more stochastic behavior patterns that are hard for participants to reason about. We expected the behaviors to elicit both positive and negative views of the robot, leading to a large variety of non-verbal reactions and impressions of robot performance.

Impressions of Robot Performance: During a navigation task, we paused the interaction at 4 seconds *before*, and at 8 seconds *after* the robot switched between behaviors. The elapsed time for the latter pause was longer in order to give people enough time to experience the latest robot behavior.

As shown in the supplementary video, impressions of robot performance were provided through an interface embedded in the simulation. The interface asked the participants to indicate their impression about the robot's most recent performance in regard to: 1) *"how competent was the robot at navigating,"* 2) *"how surprising was the robot's navigation behavior,"* and 3) *"how clear were the robot's intentions during navigation."* Participants provided ratings for these three dimensions of robot performance on a 5-point Likert responding format, e.g., with 1 being "incompetent", 2 being "somewhat incompetent", 3 being "neither competent nor incompetent", 4 being "somewhat competent", and 5 being "competent".

C. Observations

We organized observations of human-robot interactions, as recorded in SEAN-VR [73], into the features described below.

Participants' Facial Expression Features: We captured the participants' eye and lip movements, as well as their gaze through the VR headset using the VIVE Eye and Facial Tracking (SRanipal) SDK. The eye and lip movements corresponded to 73 features that described the geometry of the face through blend shapes. The gaze was a 3D vector providing the direction of gaze of the person relative to their face.

Spatial Behavior Features: During navigation, we captured the poses of the robot, the participant, and the other automatically-controlled avatars on the ground plane of the scene. Then, we computed the poses of the avatars relative to the robot, considering only those within a 7.2m radius, as this region is typically considered a robot's public space [27, 55, 33]. Each of the features were (x, y, θ) tuples with x, y being the position and θ the body orientation (yaw angle) relative to a coordinate frame attached to the robot.

Goal Features: A navigation task had an associated destination or goal that the robot had to reach. We converted the goal pose in a global frame in the warehouse to a pose in a coordinate frame attached to the robot. This pose described the robot's proximity and relative orientation to its destination.

Occupancy Features: During navigation, the robot localized [24] against a 2D map of the warehouse. We used a cropped section of the map around the robot (of 7.2m × 7.2m) to describe the occupancy of nearby space by static objects.

D. Perceived Robot Performance

Impressions of robot performance were as expected: ratings for competence and clear intention were generally higher for *Nav-Stack* than for *Spinning* and *Wrong-Way*, while the latter two tended to be more surprising than the former. Pairs of performance dimensions were significantly correlated with absolute Pearson r-values greater than 0.6. An exploratory factor analysis suggested that the dimensions could be combined into one performance factor (which explained 77% of the variance).

Using the features described before and the impressions of robot performance provided by the participants, we created a dataset of paired observation sequences and target performance values. We further refer to this data as the SEAN virTual rObot GuidE with impliciT Human fEedback and peRformance Dataset (SEAN TOGETHER Dataset). As described below, we used this dataset to investigate the questions in Sec. III.

V. Findings

A. How Well Can Human Observers Predict a User's Impression of Robot Performance?

To better understand the complexity of inferring impressions of robot performance, we evaluated how well human annotators could solve the prediction problem. To this end, we administered an online survey through `www.prolific.co`, a platform for human data collection. In the survey, human annotators observed visualizations of observations in our SEAN TOGETHER Dataset. Then, they tried to predict performance ratings provided by the people who followed the robot.

Method: For the survey, we randomly selected 2 data samples from each of the 60 participants in our data collection, with one gathered before and the other gathered after the robot's behavior changed. The observations in each sample corresponded to an 8-second 5-hz window of features right before the corresponding performance label was provided.

As shown in Fig. 3, data samples were visualized in two ways:

1. Facial Rendering. We created a human face rendering in Unity by replaying the facial expression features on an SRanipal compatible avatar, as shown in Fig. 3 (right). This visualization was motivated by the use of facial expressions in prior work on implicit feedback (e.g., [18]).

2. Navigation Rendering. We created a plot of features that described the navigation behavior of the robot and the avatars in the simulation. The plot showed features that, using existing perception techniques, may be easier to estimate than facial features in real-world deployments. These features are the spatial behavior features, the robot's goal location, the occupied space near the robot, and the gaze direction of the participant – the last of which could be approximated using an estimate of the person's head orientation [48]. Because prior work suggests that it is easier to make sense of implicit human feedback in context [12], the plot was always centered on the robot, making its surroundings always visible as in Fig. 3 (left).

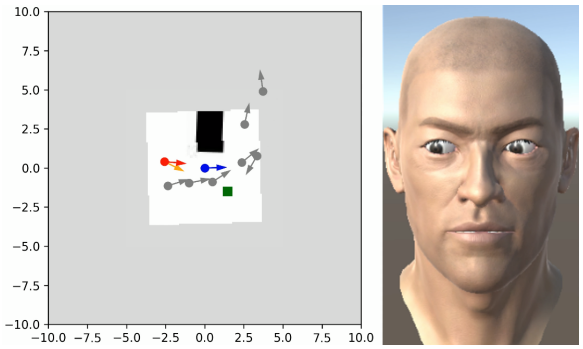

Fig. 3: A data sample from the *Nav.+Facial* condition. The **left** plot shows gaze, spatial behavior, goal, and occupancy features: ●→ is the robot's pose; ●→ is the pose of the participant following the robot during the VR interaction; → indicates the gaze of the participant; ●→ are the poses of algorithmically controlled avatars; ■ is the destination position that the robot navigated towards; and occupancy in the environment is indicated by black pixels (occupied) and white pixels (unoccupied). The **right** visualization shows a rendering of the facial expression features of the participant.

We used the visualizations to create three annotation conditions that helped understand the value of different features: 1) ***Facial-Only***: for a given data sample, annotators only saw the facial rendering; 2) ***Nav.-Only***: annotators only saw the navigation rendering; and 3) ***Nav.+Facial***: annotators saw the navigation rendering first, then the facial rendering and, finally, saw a video with both visualizations together (Fig. 3).

Each of the data samples was annotated by 10 unique people in each condition. The annotators were instructed to predict how the participant who controlled the avatar that followed the robot perceived the robot's performance. The samples they annotated were presented in random order. Each annotator was paid US$7.5 for approximately 30 min of annotation time. To encourage high-quality annotations, we also gave them a bonus of US$0.125 for each correct prediction that they made.

Annotators: We recruited a total of 100 annotators. Thirty-five of them identified as female, 60 as male, and 5 as non-binary or third gender. Ages ranged from 18 to 75 years old. Annotators indicated similar familiarity with robots (M = 4.12, SE = 0.14) as the data collection participants, though the annotators were slightly more familiar with VR (M = 4.50, SE = 0.16).

Results: We used linear mixed models estimated with REstricted Maximum Likelihood (REML) [28, 59] to analyze errors in the predictions for each performance dimension. Our independent variables were Before/After Robot Behavior Change (*Before*, *After*) and Annotation Condition (*Facial-Only*, *Nav.-Only*, *Nav.+Facial*). Also, we considered Annotator ID as a random effect because annotators provided predictions for multiple data samples. Our dependent variables were

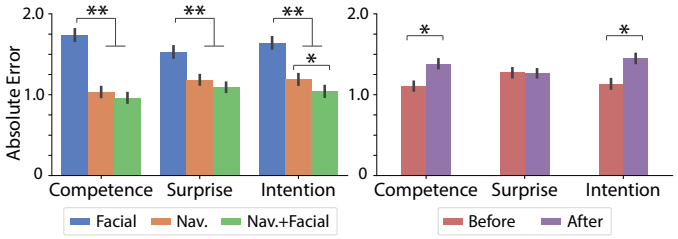

Fig. 4: Errors for annotators' predictions by (a) Annotation Conditions and (b) Before/After Robot Behavior Change. (**) and (*) denote $p < 0.0001$ and $p < 0.05$, respectively.

the absolute error between an annotator's prediction and the performance rating in our SEAN TOGETHER Dataset.

We found that the Annotation Condition had a significant effect on the absolute error for Competence, Surprise, and Intention ($p < 0.0001$ in all cases). As in Fig. 4(a), Tukey HSD post-hoc tests showed that for Competence and Surprise, the errors for *Nav.+Facial* and *Nav.-Only* were significantly lower than *Facial-Only*, yet the difference between the former two conditions was not significant. For Intention, all conditions led to significantly different errors. *Nav.+Facial* resulted in the lowest error, followed by *Nav.-Only* and then *Facial-Only*. These results suggest that facial expressions provide information about impressions of robot performance though, more generally, the features used to create the Navigation Renderings seemed to be the most critical for these predictions.

Before/After Robot Behavior Change had a significant effect on the prediction errors for Competence and Intention ($p < 0.0001$ in both cases). As in Fig. 4(b), the error was significantly lower for samples *Before* a behavior change than for samples *After* a change for these performance dimensions. We suspect this was because the robot sometimes demonstrated two behaviors in the samples collected *After* a behavior change, but the ones *Before* only showed one behavior.

Table I shows the F_1-Scores for the annotator predictions (see HA rows). The low scores suggest that correctly predicting impressions of robot performance on a 5-point responding format was difficult for humans. Despite this, we suspected that humans could do a more reasonable job at distinguishing impressions of poor robot performance from other impressions and, if this was the case, then this could open up doors in the future to using this binary signal (instead of the more fine-grained feedback) as a reward signal to adapt robot behavior in navigation tasks, e.g., in line with [37, 41]. Thus, we transformed the ground truth ratings from our data collection to binary values, one corresponding to low performance (e.g., 1-2 ratings for competence) and another to medium-to-high performance (3-5 ratings for competence). Also, we transformed the annotators' predictions similarly. This led to F_1 scores of 0.69 for Competence, 0.64 for Surprise, and 0.69 for Intention. As expected, human annotators were better at telling the directionality of robot performance ratings than at predicting their exact magnitude.

Finally, we investigated the performance of human annotations over the span of data collection because prior work suggests that the expressiveness of people engaged in human-robot interactions can change over time [13], e.g., potentially due to changes in their expectations about the robot or due to fatigue. Figures 5(a)–(c) show the evolution of mean absolute errors for the human annotators' predictions over 10-minute intervals of interaction, considering each performance dimension. In general, human performance was very stable, suggesting no major bias over time in participant's spatial behavior or facial expressions. Interestingly, the results also suggested that improvements in performance with an individual feature did not necessarily translate in improvements on the *Nav.+Facial* condition. Humans may have combined the information from the different implicit feedback modalities in subtle ways when making predictions.

B. Can Machine Learning Methods Predict Impressions of Robot Performance as Well as Humans?

We compared human prediction performance with a variety of classifiers, including a random forest and neural networks.

Method: Machine learning (ML) models were evaluated on the same samples shown to the human annotators ($n = 120$). The rest of the data was used for training ($n = 2280$) and validation ($n = 569$). We trained one model for each combination of feature sets shown to the human annotators (*Facial-Only*, *Nav.-Only*, and *Nav.+Facial*). The *Nav.* feature set included occupied space near the robot, which we encoded using a ResNet-18 representation [29]. We repeated training for each model 10 times with varying random seeds. The Random Forest (RF) used 100 trees and the depth was grown until leaves had less than 2 samples. The neural networks had a number of parameters on the same order of magnitude: 5.4×10^6 for a Multi-Layer Perceptron (MLP), 2.1×10^6 for a message-passing Graph Neural Network (GNN) [6], and 6.5×10^6 for a Transformer (T) [68]. Networks were trained using minibatch gradient descent with the Adam optimizer and cross-entropy loss. Learning rate, batch size, and dropout were chosen using grid search with validation-based early stopping [50]. We also compared all these models with a random sampling baseline.

Results: As is shown in Table I, ML models outperformed both human-level performance and random baseline in all cases when measured via F_1-Score. When measured using Accuracy and Mean Absolute Error, ML models performed the best, except for Intention when using *Nav.+Facial* features. These outcomes indicate that our implicit feedback data contained useful information that can be leveraged by ML models to predict users' impressions of robot performance. Further, ML models trained with *Nav.-Only* and *Nav.+Facial* features outperformed those trained with *Facial-Only* features. This finding aligns with our observation in Sec. V-A on the criticality of the *Nav.* features in comparison to the *Facial* features on performance prediction.

Figures 5(d)–(f) show the evolution of mean absolute errors for the Random Forest model, which generally performed the best, over 10-minute intervals of interaction during the

TABLE I: Machine learning methods and human annotation (HA) performance on 120 examples. Methods: Random (R) sampling from the distribution of labels in the training set, Random Forest (RF), Multi-Layer Perceptron (MLP), Graph Neural Network (GNN), and Transformer (T). Arrows indicate that higher (↑) and lower (↓) results are better. Cells with (-) do not have results because a GNN trained on facial features only was effectively an MLP. The Best and Second results are highlighted.

		F_1-Score ($\mu \pm \sigma$) ↑			Accuracy ($\mu \pm \sigma$) ↑			Mean Absolute Error ($\mu \pm \sigma$) ↓		
		Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial
Competence	HA	0.16 ± 0.0	0.28 ± 0.1	0.29 ± 0.2	0.19 ± 0.1	0.40 ± 0.1	0.42 ± 0.1	1.74 ± 0.2	1.03 ± 0.3	0.99 ± 0.4
	R	0.18 ± 0.0	0.19 ± 0.0	0.17 ± 0.0	0.21 ± 0.0	0.21 ± 0.0	0.20 ± 0.0	1.73 ± 0.1	1.75 ± 0.1	1.81 ± 0.1
	RF	0.19 ± 0.0	0.37 ± 0.0	0.38 ± 0.0	0.33 ± 0.0	0.52 ± 0.0	0.52 ± 0.0	1.43 ± 0.0	0.88 ± 0.0	0.82 ± 0.0
	MLP	0.23 ± 0.0	0.29 ± 0.1	0.25 ± 0.1	0.28 ± 0.0	0.48 ± 0.0	0.44 ± 0.1	1.66 ± 0.1	1.07 ± 0.3	1.19 ± 0.4
	GNN	-	0.31 ± 0.1	0.33 ± 0.0	-	0.43 ± 0.1	0.39 ± 0.1	-	1.22 ± 0.3	1.04 ± 0.0
	T	0.21 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.30 ± 0.0	0.43 ± 0.0	0.41 ± 0.1	1.58 ± 0.1	0.97 ± 0.0	0.95 ± 0.0
Surprise	HA	0.18 ± 0.0	0.24 ± 0.1	0.25 ± 0.1	0.20 ± 0.1	0.30 ± 0.1	0.32 ± 0.1	1.53 ± 0.3	1.19 ± 0.2	1.12 ± 0.2
	R	0.19 ± 0.0	0.21 ± 0.0	0.17 ± 0.0	0.20 ± 0.0	0.21 ± 0.0	0.18 ± 0.0	1.64 ± 0.1	1.60 ± 0.1	1.68 ± 0.1
	RF	0.29 ± 0.0	0.38 ± 0.0	0.34 ± 0.0	0.30 ± 0.0	0.40 ± 0.0	0.34 ± 0.0	1.30 ± 0.0	0.93 ± 0.0	0.98 ± 0.0
	MLP	0.24 ± 0.0	0.26 ± 0.1	0.24 ± 0.1	0.25 ± 0.0	0.30 ± 0.0	0.29 ± 0.1	1.23 ± 0.1	1.12 ± 0.2	1.08 ± 0.1
	GNN	-	0.29 ± 0.0	0.27 ± 0.0	-	0.30 ± 0.0	0.28 ± 0.0	-	1.13 ± 0.1	1.07 ± 0.1
	T	0.27 ± 0.0	0.29 ± 0.0	0.32 ± 0.1	0.28 ± 0.0	0.31 ± 0.0	0.33 ± 0.1	1.37 ± 0.1	1.07 ± 0.1	1.04 ± 0.1
Intention	HA	0.18 ± 0.0	0.25 ± 0.1	0.28 ± 0.1	0.21 ± 0.1	0.37 ± 0.2	0.41 ± 0.1	1.64 ± 0.2	1.19 ± 0.4	1.07 ± 0.2
	R	0.21 ± 0.1	0.19 ± 0.0	0.17 ± 0.0	0.23 ± 0.1	0.22 ± 0.0	0.19 ± 0.0	1.70 ± 0.1	1.73 ± 0.1	1.80 ± 0.1
	RF	0.28 ± 0.0	0.28 ± 0.0	0.24 ± 0.0	0.37 ± 0.0	0.43 ± 0.0	0.41 ± 0.0	1.45 ± 0.0	1.13 ± 0.0	1.14 ± 0.0
	MLP	0.27 ± 0.0	0.26 ± 0.1	0.22 ± 0.0	0.31 ± 0.0	0.41 ± 0.1	0.39 ± 0.1	1.86 ± 0.1	1.31 ± 0.3	1.51 ± 0.5
	GNN	-	0.28 ± 0.0	0.29 ± 0.0	-	0.37 ± 0.0	0.35 ± 0.0	-	1.32 ± 0.1	1.25 ± 0.1
	T	0.24 ± 0.0	0.29 ± 0.1	0.32 ± 0.0	0.33 ± 0.0	0.41 ± 0.0	0.40 ± 0.0	1.63 ± 0.1	1.21 ± 0.1	1.20 ± 0.1

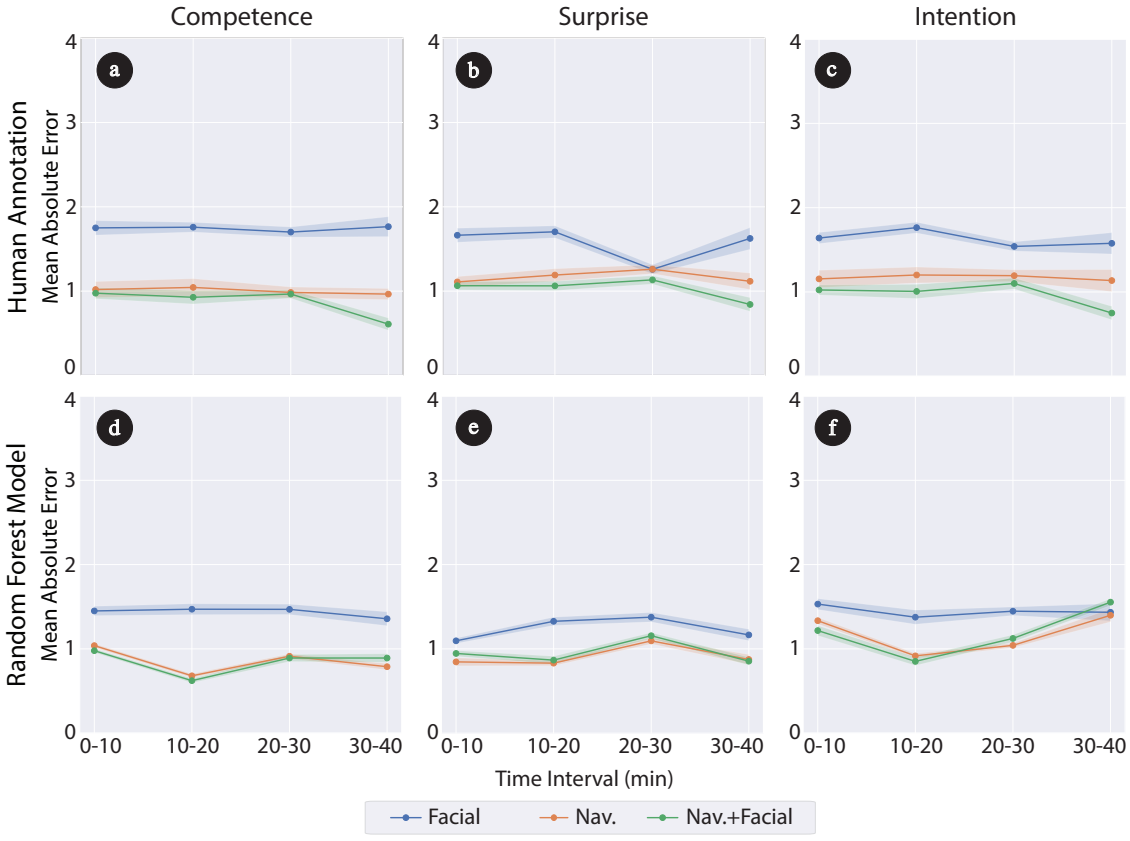

Fig. 5: Mean Absolute Errors (MAE) of human annotation and Random Forest (RF) results over 10-minute intervals of the data collection sessions. MAE was computed for all data samples in each interval, and then the average and standard errors of MAE were calculated considering the performance of the 10 unique annotators (for human annotation results in (a)–(c)) or the 10 Random Forest models trained with different seeds in Table I (RF results in (d)–(f)).

data collection. Similar to the results from human annotators (Figures 5(a)–(c), Sec. V-A), the error for the RF model did not fluctuate drastically, although the performance for Intention prediction with *Nav.* and *Nav.+Facial* features decreased in the last two time intervals of data collection (having higher mean absolute error). The decrease in performance could be the result of a distribution shift, especially in the last interval which had the fewest number of samples because not all interactions took the full 40 minutes. Also, a good proportion of the samples in the last time interval showed the end of navigation tasks, at which point the participants could have been more sensitive to robot navigation in the wrong direction. Indeed, there was a higher proportion of lower ratings for Intention in the last interval than in the other intervals, as shown in the supplementary material.

The results in Table I and Fig. 5 motivated us to focus the analysis in the next section in the aggregate, overall results rather than the interval-based results.

C. Can Machine Learning Generalize to Unseen Users?

We investigated how well learning models could predict performance by a user whose data was held out from training.

Method: We used the models and training scheme from Sec. V-B with all features (*Nav.+Facial*), but split the data using leave-one-out cross-validation. For each fold, the data for one participant was used as the test set and the remaining examples were split between training (80%) and validation (20%). We searched for new hyperparameters and computed results both on 5-classes and on binary classification. Binary targets and prediction labels were computed as in Sec. V-A.

Results: Fig. 6 reports F_1-Scores over all folds. The models generalized to unseen people with only a slight reduction in performance in comparison to Table I. Also, the average F_1-Score across all performance dimensions improves from 0.25 in the multiclass case to 0.62 in the binary case. This makes the ML predictions more usable in practice. For example, in the future, we envision deploying the trained ML on new users (as in Fig. 2b) in order to detect low robot performance. This could be an indication that the robot made a mistake, triggering interaction recovery behaviors like apologies or explanations [63], which could increase trust on the system [15].

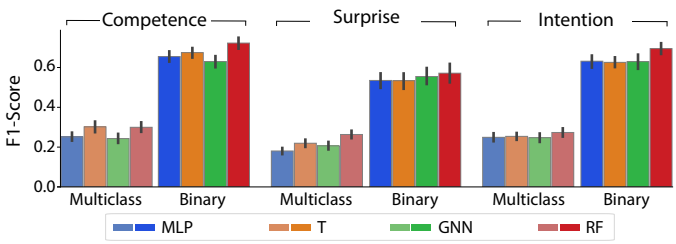

Fig. 6: ML models trained on *Nav.+Facial* features using leave-one-out cross-validation and evaluated on the held-out participant's data. F_1-Scores are computed over 5 classes (Multiclass) and 2 classes (Binary). See the text for details.

VI. Guidelines for Real-World Applications

We hope that future work leverages our findings to build effective models for mapping implicit human feedback to users' impressions of robot performance in real-world social navigation tasks. To this end, we first recommend prioritizing robust people tracking and pose estimation over computing fine-grained facial expressions, especially when computational resources may be limited. Reasoning about spatial behavior features in the context of the task can facilitate achieving reasonable prediction performance with lower sensor requirements. Also, occlusions are likely more common for facial expressions than body tracking in real-world applications.

Second, we recommend building models that focus on identifying poor robot performance (performing binary classification) instead of predicting more granular impressions of robot performance (e.g., on a 5-point scale). Even for humans, the latter type of predictions are hard because of the subjectivity of performance ratings.

Finally, if a robot is executing multiple behaviors, we recommend considering whether the robot switched behaviors recently when reasoning about performance predictions. As in our results, predicting performance recently after a behavior change can be more difficult than before, when the behavior was more consistent.

VII. Limitations and Future Work

Our work has several limitations that point to interesting future directions. First, we obtained human baselines for prediction performance, but used only a limited set of feature combinations. In the future, it would be interesting to consider a broader set of feature categories. For instance, future work could evaluate the individual impact of gaze versus other facial features, and further study the value of considering more detailed human pose features (e.g., [74]) when inferring robot performance. Second, because participants were wearing Vive Pro Eye VR headsets in our data collection, our study was not able to capture images of their real face. We instead used a rendered face to visualize the captured eye and lip features (as in Fig. 3) and 73 features provided by the headset to describe the geometry of the face through blend shapes. However, this specific visualization and featurization could have lost details from subtle expressions that could be useful to predict robot performance. In the future, it would be interesting to utilize more advanced devices such as the recently released Apple Vision Pro to create other datasets of implicit human feedback. The new Apple device can sense faces in a way that allows rendering higher quality avatars for users, and the data it captures could potentially improve the accuracy and robustness of ML models that predict robot performance. Third, our work focused on navigation in a VR setup. An immediate next step is to extend our work to real-world interactions, verifying the generalizability of prediction models to different tasks and considering sensor noise in the detected features. Lastly, the inferred performance predictions, which could be considered instantaneous rewards, could be used in the future to adapt robot behavior in HRI [37, 41, 18].

VIII. Conclusion

This work contributes the SEAN TOGETHER Dataset, consisting of observations of human-robot interactions in VR, including implicit human feedback, and corresponding performance ratings in guided robot navigation tasks. Our analyses revealed that facial expressions can help predict impressions of the robot, but spatial behavior features in the context of the navigation task were more critical for these inferences. Our experiments also demonstrated the ability of humans and ML models to infer perceived robot performance from interaction observations. Also, ML models were better at predicting the directionality of impressions of robot performance (as a binary classification task) than predicting exact performance ratings (on a 5-point scale). Importantly, the models more than doubled in F_1-Score performance with binary classification than multi-class classification, showing good potential to generalize to novel users in the former case. Our dataset, accompanying analyses, and guidelines pave a path forward to enabling mobile robots to leverage passive observations of their users to infer how well they complete navigation tasks. Potentially, they could also use this feedback to interactively improve their behavior in the future.

Acknowledgments

Removed for blind review.

References

[1] Georgios Angelopoulos, Alessandra Rossi, Claudia Di Napoli, and Silvia Rossi. You are in my way: Non-verbal social cues for legible robot navigation behaviors. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 657–662. IEEE, 2022.

[2] Reuben M Aronson and Henny Admoni. Gaze for error detection during human-robot shared manipulation. In *Fundamentals of Joint Action workshop, Robotics: Science and Systems*, page 5, 2018.

[3] Chatchalita Asavanant and Hiroyuki Umemuro. Personal space violation by a robot: An application of expectation violation theory in human-robot interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 1181–1188. IEEE, 2021.

[4] Eleanor Avrunin and Reid Simmons. Socially-appropriate approach paths using human data. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1037–1042. IEEE, 2014.

[5] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. *Human-robot interaction: An introduction*. Cambridge University Press, 2020.

[6] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[7] Aniket Bera, Tanmay Randhavane, and Dinesh Manocha. Improving socially-aware multi-channel human emotion prediction for robot navigation. In *CVPR Workshops*, pages 21–27, 2019.

[8] Erdem Bıyık, Aditi Talati, and Dorsa Sadigh. Aprel: A library for active preference-based reward learning algorithms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 613–617. IEEE, 2022.

[9] Rita Borgo, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, Fintan McGee, Luana Micallef, Tatiana von Landesberger, et al. Crowdsourcing for information visualization: Promises and pitfalls. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22–27, 2015, Revised Contributions*, pages 96–138. Springer, 2017.

[10] Martim Brandao, Gerard Canal, Senka Krivić, Paul Luff, and Amanda Coles. How experts explain motion planner output: a preliminary user-study to inform the design of explainable planners. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 299–306. IEEE, 2021.

[11] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1): 82–111, 2013.

[12] Kate Candon, Jesse Chen, Yoony Kim, Zoe Hsu, Nathan Tsoi, , and Marynel Vázquez. Nonverbal human signals can help autonomous agents infer human preferences for their behavior. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[13] Kate Candon, Nicholas C. Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. React: Two datasets for analyzing both human reactions and evaluative feedback to robots over time, 2024.

[14] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. The robotic social attributes scale (rosas) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, pages 254–262, 2017.

[15] Yuhang Che, Allison M Okamura, and Dorsa Sadigh. Efficient and trustworthy social navigation via explicit and implicit robot–human communication. *IEEE Transactions on Robotics*, 36(3):692–707, 2020.

[16] Mohamed Chetouani. Interactive robot learning: An overview. *ECCAI Advanced Course on Artificial Intelligence*, pages 140–172, 2021.

[17] Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca

Fitzgerald. Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In *International Joint Conference on Artificial Intelligence*, 2021.

[18] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, pages 604–626. PMLR, 2021.

[19] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[20] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 51–58, 2015.

[21] Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740*, 2023.

[22] Yuxiang Gao and Chien-Ming Huang. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI*, 8:721317, 2022.

[23] Rachel Gockley, Jodi Forlizzi, and Reid Simmons. Natural person-following behavior for social robots. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 17–24, 2007.

[24] Giorgio Grisetti, Cyrill Stachniss, and Wolfram Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE transactions on Robotics*, 23(1):34–46, 2007.

[25] Balint Gucsi, Danesh S Tarapore, William Yeoh, Christopher Amato, and Long Tran-Thanh. To ask or not to ask: A user annoyance aware preference elicitation framework for social robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7935–7940. IEEE, 2020.

[26] Hatice Gunes, Massimo Piccardi, and Maja Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In *Affective Computing*. IntechOpen, 2008.

[27] Edmund T Hall and Edward T Hall. *The hidden dimension*, volume 609. Anchor, 1966.

[28] David A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358): 320–338, 1977. ISSN 01621459. URL http://www.jstor.org/stable/2286796.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Tom P Huck, Christoph Ledermann, and Torsten Kröger. Testing robot system safety by creating hazardous human worker behavior in simulation. *IEEE Robotics and Automation Letters*, 7(2):770–777, 2021.

[31] Angel Hsing-Chi Hwang and Andrea Stevenson Won. Ideabot: investigating social facilitation in human-machine team creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[32] Tetsunari Inamura, Yoshiaki Mizuchi, and Hiroki Yamada. Vr platform enabling crowdsourcing of embodied hri experiments–case study of online robot competition. *Advanced Robotics*, 35(11):697–703, 2021.

[33] Walther Jensen, Simon Hansen, and Hendrik Knoche. Knowing you, seeing me: Investigating user preferences in drone-human acknowledgement. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[34] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022.

[35] Adam Kendon. Goffman's approach to face-to-face interaction. *Erving Goffman: Exploring the interaction order*, 1988.

[36] Ross A Knepper, Christoforos I Mavrogiannis, Julia Proft, and Claire Liang. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, pages 283–292, 2017.

[37] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

[38] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765, 2020.

[39] Shih-Yun Lo, Katsu Yamane, and Ken-ichiro Sugiyama. Perception of pedestrian avoidance strategies of a self-balancing mobile robot. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1243–1250. IEEE, 2019.

[40] David V Lu, Dave Hershberger, and William D Smart. Layered costmaps for context-sensitive navigation. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 709–715. IEEE, 2014.

[41] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, pages 2285–2294. PMLR, 2017.

[42] Roberto Martín-Martín, Hamid Rezatofighi, Abhijeet

Shenoi, Mihir Patel, J Gwak, Nathan Dass, Alan Federman, Patrick Goebel, and Silvio Savarese. Jrdb: A dataset and benchmark for visual perception for navigation in human environments. *arXiv preprint arXiv:1910.11792*, 2019.

[43] Christoforos Mavrogiannis, Patrícia Alves-Oliveira, Wil Thomason, and Ross A Knepper. Social momentum: Design and evaluation of a framework for socially competent robot navigation. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(2):1–37, 2022.

[44] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):1–39, 2023.

[45] Emily McQuillin, Nikhil Churamani, and Hatice Gunes. Learning socially appropriate robo-waiter behaviours through real-time user feedback. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 541–550. IEEE, 2022.

[46] Daxton Mitchell, HeeSun Choi, and Justin M Haney. Safety perception and behaviors during human-robot interaction in virtual environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 2087–2091. SAGE Publications Sage CA: Los Angeles, CA, 2020.

[47] Noriaki Mitsunaga, Christian Smith, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Adapting robot behavior for human–robot interaction. *IEEE Transactions on Robotics*, 24(4):911–916, 2008.

[48] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE, 2016.

[49] Sören Pirk, Edward Lee, Xuesu Xiao, Leila Takayama, Anthony Francis, and Alexander Toshev. A protocol for validating social navigation policies. *arXiv preprint arXiv:2204.05443*, 2022.

[50] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[51] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. volume 3, 01 2009.

[52] Claire Rivoire and Angelica Lim. The delicate balance of boring and annoying: Learning proactive timing in long-term human robot interaction. 2016.

[53] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and systems*, volume 2, pages 1–9. Ann Arbor, MI, USA, 2016.

[54] Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29, 2018.

[55] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. A larger audience, please!—encouraging people to listen to a guide robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 31–38. IEEE, 2010.

[56] Maia Stiber. Effective human-robot collaboration via generalized robot error management using natural human responses. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 673–678, 2022.

[57] Maia Stiber, Russell Taylor, and Chien-Ming Huang. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 676–683. IEEE, 2022.

[58] Maia Stiber, Russell H. Taylor, and Chien-Ming Huang. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 222–230, New York, NY, USA, 2023. Association for Computing Machinery. URL https://doi.org/10.1145/3568162.3576990.

[59] Walter W Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

[60] Aamodh Suresh, Angelique Taylor, Laurel D Riek, and Sonia Martinez. Robot navigation in risky, crowded environments: Understanding human preferences. *arXiv preprint arXiv:2303.08284*, 2023.

[61] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 114–122. IEEE, 2019.

[62] Andrea L Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172 (6-7):716–737, 2008.

[63] Leimin Tian and Sharon Oviatt. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2):1–32, 2021.

[64] Russell Toris, David Kent, and Sonia Chernova. The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing. *Journal of Human-Robot Interaction*, 3(2):25–49, 2014.

[65] Pete Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. *The International Journal of Robotics Research*, 34(3):335–356, 2015.

[66] Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa,

JD Zhao, and Marynel Vázquez. An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7528–7535. IEEE, 2021.

[67] Nathan Tsoi, Alec Xiang, Peter Yu, Samuel S Sohn, Greg Schwartz, Subashri Ramesh, Mohamed Hussein, Anjali W Gupta, Mubbasir Kapadia, and Marynel Vázquez. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters*, 7(4):11047–11054, 2022.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[69] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[70] Lennart Wachowiak, Peter Tisnikar, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. Analysing eye gaze patterns during confusion and errors in human–agent collaborations. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 224–229. IEEE, 2022.

[71] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. Frownonerror: Interrupting responses from smart speakers by facial expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[72] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. Self-annotation methods for aligning implicit and explicit human feedback in human-robot interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 398–407, 2023.

[73] Qiping Zhang, Nathan Tsoi, and Marynel Vázquez. Sean-vr: An immersive virtual reality experience for evaluating social robot navigation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 902–904, 2023.

[74] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019.