Towards Inferring Users' Impressions of Robot Performance in Navigation Scenarios

Qiping Zhang[1*], Nathan Tsoi[1*], Booyeon Choi[1], Jie Tan[2], Hao-Tien Lewis Chiang[2], Marynel Vázquez[1]

***Abstract*—Human impressions of robot performance are often measured through surveys. As a more scalable and cost-effective alternative, we study 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 combinations of observation types (e.g., facial, spatial, and map features). Our results show that facial expressions alone provide useful information about human impressions of robot performance; but in the navigation scenarios we tested, spatial features are the most critical piece of information for this inference task. Also, when evaluating results as binary classification (rather than multiclass classification), the F_1-Score of human predictions and machine learning models more than doubles, showing that both are better at telling the directionality of robot performance than predicting exact performance ratings. Based on our findings, we provide guidelines for implementing these predictions models 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 [1]–[4]. The feedback corresponds to communicative signals that are inevitably given off by people [5]. They can be reflected in human actions that change the world's physical state [6] or can be nonverbal cues, such as facial expressions [2], [3] and gaze [1], [7], displayed during social interactions. Implicit feedback serves as a burden-free information channel that sometimes persists even when people don't intend to communicate [8].

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 [9] or robot photography [4], 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. 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 controlled laboratory 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.

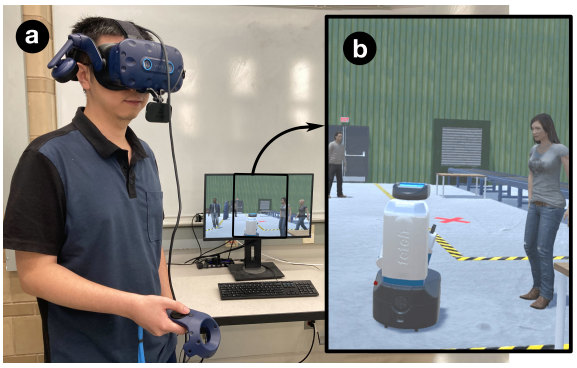

Fig. 1: Data collection. Humans controlled an avatar in the simulation with VR (a) while they were guided by a Fetch robot (b). The screen on the desk shows what the user saw.

Due to the complexity of reliably capturing observations of implicit feedback during navigation tasks, we performed a data collection effort using the Social Environment for Autonomous Navigation (SEAN) 2.0 [10] with Virtual Reality (VR) [11].[3] 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 [4], because humans may care about many aspects of a robot's navigation behavior, as discussed in the social robot navigation literature [12]–[14].

Using the data collected with SEAN, we first investigate to what extent humans can predict the users' impression of a robot's performance (along the dimensions of perceived competence, surprise, and intention) from a visualization of the observations of interactions (or features) recorded in our

This work was supported by the National Science Foundation (Grant No. IIS-1924802 and IIS-2143109) and Google. We are grateful to Carolina Parada and Leila Takayama for their valuable feedback.

[1]Qiping Zhang, Nathan Tsoi, Booyeon Choi, and Marynel Vázquez are with Yale University. [2]Jie Tan and Lewis Chiang are with Google DeepMind.
*Equal contribution. Corresponding author: qiping.zhang@yale.edu.

[3]Dataset available at: https://sean-together.interactive-machines.com/

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 impressions of robot performance in navigation tasks and the value of various combinations of features in this inference problem. 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.

II. Related Work

Impressions of Robot Performance. Understanding human impressions of robot performance is important. They can be used to evaluate robot policies [15]–[17] and to create better robot behavior [7], [18]–[20], increasing the likelihood of robot adoption. In this work, we focus on inferring three robot performance dimensions relevant to navigation [12]: robot competence, the surprisingness of robot behavior, and clear intent. Robot competence is a popular performance metric [21], especially in robot navigation [22]–[24]. Surprising behavior violates expectations. It is often considered undesired [14], [25] and may require explanations by the robot [26]. Meanwhile, showing clear intent means that the robot enables an observer to infer the goal of its motion [27]. 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 [28], [29].

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 [30], [31] or survey instruments [22], [32]. 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 [33], [34], thereby minimizing the risk of feedback fatigue [35]. Learning from implicit feedback is not without challenges, however, as it can be difficult to interpret [2], [3]. For example, this can happen due to inter-person variability in facial expressions [36] or similar signals being produced for different reasons [37].

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 [38]. While in some cases these signals are explicit feedback (e.g., to interrupt an agent [39]), our work considers them implicit feedback because we do not prime humans to react in specific ways to a robot. As such, our work is closer to [2], [37], [40]–[42], which used nonverbal signals to identify critical states during robot operation, detect robot errors, and adjust robot behavior. Other types of feedback signals, such as those from brain-computer interfaces, have been used in HRI [43]–[45]; however, they are impractical for navigation tasks.

Simulation in HRI. Simulation is a useful tool in HRI [46]–[48], and particularly popular in robot navigation research [23], [49]–[51]. Robotics simulators model aspects of the real world in a virtual environment and render virtual representations of the real world, often using game engines such as Unity [10], [52], [53]. In this work, we take advantage of the SEAN 2.0 simulator [10], which integrates with the Robot Operating System, and supports VR [11]. Virtual Reality interfaces have gained popularity in HRI [54]–[59]. Some VR systems, such as the Vive Pro Eye [60] that we utilize in our work (Fig. 1), also allow tracking of eye-gaze and facial features.

Human Annotations. We build on prior HRI research that utilizes user self-reports (or self-annotations) to create prediction models relevant to a task of interest [4], [61]. Self-reports consist of first-hand opinions from users about their experiences [62]. In HRI, these are opinions by direct users of robots – rather than opinions by third-parties that observe the experiences [23], [63]–[66]. For instance, [4] asked study participants to evaluate robot performance using video logs immediately after they interacted with a robot. Similarly, we asked robot users to evaluate robot performance. However, instead of discretizing interactions based on high-level robot actions and collecting impressions of robot performance all throughout interactions, we opted for querying humans about their impressions of the robot at critical points in time during a navigation task. This was necessary because navigation actions are continuous, rather than discrete as in [4]. This makes it very time-consuming and expensive to both segment robot behavior and annotate performance across whole interactions.

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. Specifically, we aim to learn a mapping from a sequence of observations to an individual's reported impressions at the end of the sequence. We consider multiple robot performance dimensions on a 5-point scale, as detailed later in Sec. IV.

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 main 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 at a given time-step 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 [2], [4], [9], [37]. 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. 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

We collected data using SEAN-VR [11]. As in Fig. 1(a), participants used a Vive Pro Eye VR device to control an avatar in a warehouse. They had to follow a Fetch robot that guided them to a destination that was unknown to them a priori. The VR headset captured implicit signals from the participants, like eye and lip movements. Also, participants provided ratings of robot performance through the simulation's VR interface.

Fig. 1(b) shows an example first-person view of the simulation during robot-guided navigation. The Fetch robot was controlled with the Robot Operating System (ROS) [67] in SEAN. The environment contained other algorithmically controlled pedestrians and obstacles typical of warehouses. 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. 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 behaviors:

1. Nav-Stack. The robot navigated efficiently to the destination based on the path planned by the ROS Navigation Stack with social costs [68]. This behavior lasted 40 seconds.

2. Spinning. The robot rotated at its current position, indicating confusion. This behavior lasted 20 seconds.

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.

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. We expected the behaviors to elicit both positive and negative views of the robot.

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 [11], 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 that were up to 7.2m away from it, as this region is typically considered a robot's public space [69]–[71]. Each of the features were (x, y, θ) tuples with x, y being the position and θ being 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 [72] against a 2D map of the warehouse. We used a cropped section of the map around the robot (of 7.2m $\times$ 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. Analyses

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.

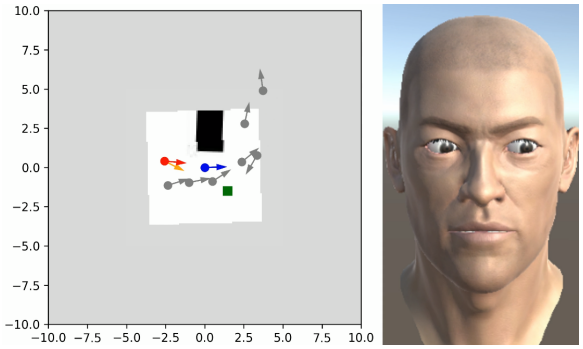

Fig. 2: A data sample from the *Nav.+Facial* condition. The **left** plot shows gaze, spatial behavior, goal, and occupancy features: ●→ is the robot's pose; ●→ is the pose of the participant following the robot during the VR interaction; → indicates the gaze of the participant; ●→ are the poses of algorithmically 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.

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. 2, 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. 2 (right). This visualization was motivated by the use of facial expressions in prior work on implicit feedback (e.g., [2]).

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 [73]. Because prior work suggests that it is easier to make sense of implicit human feedback in context [37], the plot was always centered on the robot, making its surroundings always visible as in Fig. 2 (left).

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. 2).

Each of the data samples was annotated by 10 unique people in each condition. The annotators were instructed to predict

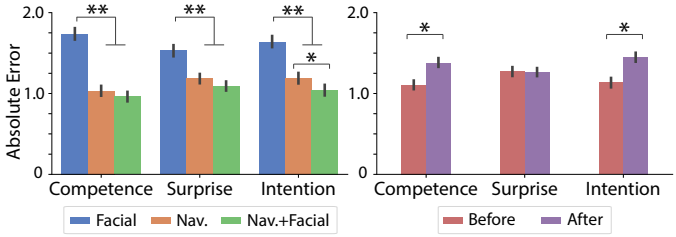

Fig. 3: Errors for annotators' predictions by (a) Annotation Conditions and (b) Before/After Robot Behavior Change. (**) and (*) denote $p < 0.0001$ and $p < 0.05$, respectively.

how the participant who controlled the avatar to follow the robot perceived the robot's performance. 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-two of them identified as female, 61 as male, and 7 as non-binary or third gender. Ages ranged from 18 to 76 years old. Annotators indicated similar familiarity with robots (M = 4.13, SE = 0.14) as the data collection participants, though the annotators were slightly more familiar with VR (M = 4.07, SE = 0.17).

Results: We used linear mixed models estimated with REstricted Maximum Likelihood (REML) [74], [75] 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 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. 3(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 seem 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. 3(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 2 behaviors in the samples collected *After* a behavior change.

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. To better understand annotators' predictions, 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, suggesting that human annotators were better at telling the directionality of robot performance ratings than at predicting their exact magnitude.

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 ($n = 569$). One model was trained 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 [76]. 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) [77], and 6.5×10^6 for a Transformer (T) [78]. 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 [79]. 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 contain 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 result aligns with our observation in Sec. V-A on the criticality of Navigation features in comparison to Facial Expressions.

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

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

		F_1-Score ↑			Accuracy ↑			Mean Absolute Error ↓		
		Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial
Competence	HA	0.16 ± 0.0	0.28 ± 0.1	0.30 ± 0.1	0.19 ± 0.1	0.40 ± 0.1	0.43 ± 0.1	1.74 ± 0.2	1.03 ± 0.3	0.96 ± 0.3
	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	$\mathbf{0.37 \pm 0.0}$	$\mathbf{0.38 \pm 0.0}$	$\mathbf{0.33 \pm 0.0}$	$\mathbf{0.52 \pm 0.0}$	$\mathbf{0.52 \pm 0.0}$	$\mathbf{1.43 \pm 0.0}$	$\mathbf{0.88 \pm 0.0}$	$\mathbf{0.82 \pm 0.0}$
	MLP	$\mathbf{0.23 \pm 0.0}$	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	0.33 ± 0.1	1.53 ± 0.3	1.19 ± 0.2	1.09 ± 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	$\mathbf{0.29 \pm 0.0}$	$\mathbf{0.38 \pm 0.0}$	$\mathbf{0.34 \pm 0.0}$	$\mathbf{0.30 \pm 0.0}$	$\mathbf{0.40 \pm 0.0}$	$\mathbf{0.34 \pm 0.0}$	1.30 ± 0.0	$\mathbf{0.93 \pm 0.0}$	$\mathbf{0.98 \pm 0.0}$
	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	$\mathbf{1.23 \pm 0.1}$	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	0.30 ± 0.1	0.21 ± 0.1	0.37 ± 0.2	$\mathbf{0.42 \pm 0.1}$	1.64 ± 0.2	1.19 ± 0.4	$\mathbf{1.04 \pm 0.3}$
	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	$\mathbf{0.28 \pm 0.0}$	0.28 ± 0.0	0.24 ± 0.0	$\mathbf{0.37 \pm 0.0}$	$\mathbf{0.43 \pm 0.0}$	0.41 ± 0.0	$\mathbf{1.45 \pm 0.0}$	$\mathbf{1.13 \pm 0.0}$	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	$\mathbf{0.29 \pm 0.1}$	$\mathbf{0.32 \pm 0.0}$	0.33 ± 0.0	0.41 ± 0.0	0.40 ± 0.0	1.63 ± 0.1	1.21 ± 0.1	1.20 ± 0.1

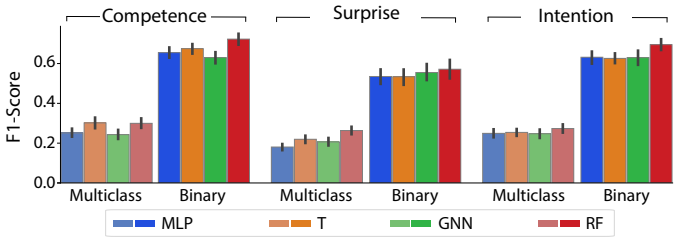

Fig. 4: ML models trained on *Nav.+Facial* features using leave-one-out cross-validation and evaluated on the held-out participant's data. F_1-Scores are computed over 5 classes (Multiclass) and 2 classes (Binary). See the text for details.

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. 4 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 begins to make the ML predictions usable for real-world applications.

VI. Discussion

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. Second, we recommend building models that focus on identifying poor robot performance instead of predicting more specific 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.

Limitations and Future Work: Our work has several limitations. 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. Second, 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.

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 dataset and accompanying analyses 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.

REFERENCES

[1] R. M. Aronson and H. Admoni, "Gaze for error detection during human-robot shared manipulation," in *Fundamentals of Joint Action workshop, Robotics: Science and Systems*, 2018, p. 5.

[2] Y. Cui, Q. Zhang, B. Knox, A. Allievi, P. Stone, and S. Niekum, "The empathic framework for task learning from implicit human feedback," in *Conference on Robot Learning*. PMLR, 2021, pp. 604–626.

[3] M. Stiber, R. Taylor, and C.-M. Huang, "Modeling human response to robot errors for timely error detection," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 676–683.

[4] Q. Zhang, A. Narcomey, K. Candon, and M. 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*, 2023, pp. 398–407.

[5] R. A. Knepper, C. I. Mavrogiannis, J. Proft, and C. Liang, "Implicit communication in a joint action," in *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, 2017, pp. 283–292.

[6] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for autonomous cars that leverage effects on human actions." in *Robotics: Science and systems*, vol. 2. Ann Arbor, MI, USA, 2016, pp. 1–9.

[7] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Adapting robot behavior for human–robot interaction," *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 911–916, 2008.

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

[9] M. Stiber, R. H. Taylor, and C.-M. Huang, "On using social signals to enable flexible error-aware hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 222–230. [Online]. Available: https://doi.org/10.1145/3568162.3576990

[10] N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez, "Sean 2.0: Formalizing and generating social situations for robot navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 047–11 054, 2022.

[11] Q. Zhang, N. Tsoi, and M. 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*, 2023, pp. 902–904.

[12] Y. Gao and C.-M. Huang, "Evaluation of socially-aware robot navigation," *Frontiers in Robotics and AI*, vol. 8, p. 721317, 2022.

[13] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[14] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *arXiv preprint arXiv:2306.16740*, 2023.

[15] X. Z. Tan, S. Reig, E. J. Carter, and A. 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)*. IEEE, 2019, pp. 114–122.

[16] S.-Y. Lo, K. Yamane, and K.-i. Sugiyama, "Perception of pedestrian avoidance strategies of a self-balancing mobile robot," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1243–1250.

[17] S. Pirk, E. Lee, X. Xiao, L. Takayama, A. Francis, and A. Toshev, "A protocol for validating social navigation policies," *arXiv preprint arXiv:2204.05443*, 2022.

[18] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.

[19] Y. Cui, P. Koppol, H. Admoni, S. Niekum, R. Simmons, A. Steinfeld, and T. Fitzgerald, "Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning," in *International Joint Conference on Artificial Intelligence*, 2021.

[20] A. Bera, T. Randhavane, and D. Manocha, "Improving socially-aware multi-channel human emotion prediction for robot navigation." in *CVPR Workshops*, 2019, pp. 21–27.

[21] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

[22] C. Mavrogiannis, P. Alves-Oliveira, W. Thomason, and R. A. Knepper, "Social momentum: Design and evaluation of a framework for socially competent robot navigation," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 2, pp. 1–37, 2022.

[23] N. Tsoi, M. Hussein, O. Fugikawa, J. Zhao, and M. 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)*. IEEE, 2021, pp. 7528–7535.

[24] G. Angelopoulos, A. Rossi, C. Di Napoli, and S. 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)*. IEEE, 2022, pp. 657–662.

[25] C. Asavanant and H. 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)*. IEEE, 2021, pp. 1181–1188.

[26] M. Brandao, G. Canal, S. Krivić, P. Luff, and A. 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)*. IEEE, 2021, pp. 299–306.

[27] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[28] A. Sciutti, M. Mara, V. Tagliasco, and G. Sandini, "Humanizing human-robot interaction: On the importance of mutual understanding," *IEEE Technology and Society Magazine*, vol. 37, no. 1, pp. 22–29, 2018.

[29] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 51–58.

[30] E. Bıyık, A. Talati, and D. Sadigh, "Aprel: A library for active preference-based reward learning algorithms," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 613–617.

[31] A. Suresh, A. Taylor, L. D. Riek, and S. Martinez, "Robot navigation in risky, crowded environments: Understanding human preferences," *arXiv preprint arXiv:2303.08284*, 2023.

[32] E. Avrunin and R. Simmons, "Socially-appropriate approach paths using human data," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 1037–1042.

[33] C. Rivoire and A. Lim, "The delicate balance of boring and annoying: Learning proactive timing in long-term human robot interaction," 2016.

[34] B. Gucsi, D. S. Tarapore, W. Yeoh, C. Amato, and L. 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)*. IEEE, 2020, pp. 7935–7940.

[35] J. Lin, Z. Ma, R. Gomez, K. Nakamura, B. He, and G. Li, "A review on interactive reinforcement learning from human social feedback," *IEEE Access*, vol. 8, pp. 120 757–120 765, 2020.

[36] H. Gunes, M. Piccardi, and M. Pantic, "From the lab to the real world: Affect recognition using multiple cues and modalities," in *Affective Computing*. IntechOpen, 2008.

[37] 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 *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[38] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[39] Y. Yan, C. Yu, W. Zheng, R. Tang, X. Xu, and Y. Shi, "Frownonerror: Interrupting responses from smart speakers by facial expressions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[40] L. Wachowiak, P. Tisnikar, G. Canal, A. Coles, M. Leonetti, and O. 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)*. IEEE, 2022, pp. 224–229.

[41] E. McQuillin, N. Churamani, and H. Gunes, "Learning socially appropriate robo-waiter behaviours through real-time user feedback," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 541–550.

[42] M. Stiber, "Effective human-robot collaboration via generalized robot error management using natural human responses," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 673–678.

[43] D. P.-O. Bos, B. Reuderink, B. van de Laar, H. Gürkök, C. Mühl, M. Poel, D. Heylen, and A. Nijholt, "Human-computer interaction for bci games: Usability and user experience," in *2010 International Conference on Cyberworlds*. IEEE, 2010, pp. 277–281.

[44] K. Muelling, A. Venkatraman, J.-S. Valois, J. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, "Autonomy infused teleoperation with application to bci manipulation," *arXiv preprint arXiv:1503.05451*, 2015.

[45] D. Xu, M. Agarwal, E. Gupta, F. Fekri, and R. Sivakumar, "Accelerating reinforcement learning agent with eeg-based implicit human feedback," *arXiv preprint arXiv:2006.16498*, 2020.

[46] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard: simulating the human for interaction research," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 101–108.

[47] S. Lemaignan, M. Hanheide, M. Karg, H. Khambhaita, L. Kunze, F. Lier, I. Lütkebohle, and G. Milliez, "Simulation and hri recent perspectives with the morse simulator," in *Simulation, Modeling, and Programming for Autonomous Robots: 4th International Conference, SIMPAR 2014, Bergamo, Italy, October 20-23, 2014. Proceedings 4*. Springer, 2014, pp. 13–24.

[48] G. Silvera, A. Biswas, and H. Admoni, "Dreye vr: Democratizing virtual reality driving simulation for behavioural & interaction research," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 639–643.

[49] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain *et al.*, "igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," *arXiv preprint arXiv:2108.03272*, 2021.

[50] A. Favier, P.-T. Singamaneni, and R. Alami, "An intelligent human simulation (inhus) for developing and experimenting human-aware and interactive robot abilities," 2021.

[51] J. Hart, R. Mirsky, X. Xiao, and P. Stone, "Incorporating gaze into social navigation," *arXiv preprint arXiv:2107.04001*, 2021.

[52] L. Kästner, T. Bhuiyan, T. A. Le, E. Treis, J. Cox, B. Meinardus, J. Kmiecik, R. Carstens, D. Pichel, B. Fatloun *et al.*, "Arena-bench: A benchmarking suite for obstacle avoidance approaches in highly dynamic environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9477–9484, 2022.

[53] Y. Liu, G. Novotny, N. Smirnov, W. Morales-Alvarez, and C. Olaverri-Monreal, "Mobile delivery robots: mixed reality-based simulation relying on ros and unity 3d," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 15–20.

[54] T. Williams, D. Szafir, T. Chakraborti, and H. Ben Amor, "Virtual, augmented, and mixed reality for human-robot interaction," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 403–404.

[55] T. Inamura and Y. Mizuchi, "Sigverse: A cloud-based vr platform for research on multimodal human-robot interaction," *Frontiers in Robotics and AI*, vol. 8, p. 549360, 2021.

[56] M. Dianatfar, J. Latokartano, and M. Lanz, "Review on existing vr/ar solutions in human–robot collaboration," *Procedia CIRP*, vol. 97, pp. 407–411, 2021.

[57] R. Suzuki, A. Karim, T. Xia, H. Hedayati, and N. Marquardt, "Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–33.

[58] O. Phaijit, M. Obaid, C. Sammut, and W. Johal, "A taxonomy of functional augmented reality for human-robot interaction," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 294–303.

[59] M. Walker, T. Phung, T. Chakraborti, T. Williams, and D. Szafir, "Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy," *arXiv preprint arXiv:2202.11249*, 2022.

[60] VIVE, "Vive pro eye overview," 2019. [Online]. Available: https://www.vive.com/sea/product/vive-pro-eye/overview/

[61] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, 2019.

[62] M. F. Jung, "Coupling interactions and performance: Predicting team performance from thin slices of conflict," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 23, no. 3, pp. 1–32, 2016.

[63] N. T. Fitter and K. J. Kuchenbecker, "Designing and assessing expressive open-source faces for the baxter robot," in *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*. Springer, 2016, pp. 340–350.

[64] M. A. Rana, D. Chen, J. Williams, V. Chu, S. R. Ahmadzadeh, and S. Chernova, "Benchmark for skill learning from demonstration: Impact of user experience, task complexity, and start configuration on performance," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7561–7567.

[65] E. Cha, N. T. Fitter, Y. Kim, T. Fong, and M. Matarić, "Generating expressive light signals for appearance-constrained robots," in *Proceedings of the 2018 International Symposium on Experimental Robotics*. Springer, 2020, pp. 595–607.

[66] P. Jonell, Y. Yoon, P. Wolfert, T. Kucherenko, and G. E. Henter, "Hemvip: Human evaluation of multiple videos in parallel," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 707–711.

[67] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," vol. 3, 01 2009.

[68] D. V. Lu, D. Hershberger, and W. D. Smart, "Layered costmaps for context-sensitive navigation," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 709–715.

[69] E. T. Hall and E. T. Hall, *The hidden dimension*. Anchor, 1966, vol. 609.

[70] M. Shiomi, T. Kanda, H. Ishiguro, and N. 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)*. IEEE, 2010, pp. 31–38.

[71] W. Jensen, S. Hansen, and H. 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*, 2018, pp. 1–12.

[72] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.

[73] O. Palinko, F. Rea, G. Sandini, and A. 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)*. IEEE, 2016, pp. 5048–5054.

[74] D. A. Harville, "Maximum likelihood approaches to variance component estimation and to related problems," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 320–338, 1977. [Online]. Available: http://www.jstor.org/stable/2286796

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

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[77] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[79] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*. Springer, 2002, pp. 55–69.