Learning Gaze Behaviors for Balancing Participation in Group Human-Robot Interactions

Sarah Gillet, Maria Teresa Parreira
KTH Royal Institute of Technology
sgillet@kth.se, parreira@kth.se

Marynel Vázquez
Yale University
marynel.vazquez@yale.edu

Iolanda Leite
KTH Royal Institute of Technology
iolanda@kth.se

***Abstract*—Robots can affect group dynamics. In particular, prior work has shown that robots that use hand-crafted gaze heuristics can influence human participation in group interactions. However, hand-crafting robot behaviors can be difficult and might have unexpected results in groups. Thus, this work explores learning robot gaze behaviors that balance human participation in conversational interactions. More specifically, we examine two techniques for learning a gaze policy from data: imitation learning (IL) and batch reinforcement learning (RL). First, we formulate the problem of learning a gaze policy as a sequential decision-making task focused on human turn-taking. Second, we experimentally show that IL can be used to combine strategies from hand-crafted gaze behaviors, and we formulate a novel reward function to achieve a similar result using batch RL. Finally, we conduct an offline evaluation of IL and RL policies and compare them via a user study (N=50). The results from the study show that the learned behavior policies did not compromise the interaction. Interestingly, the proposed reward for the RL formulation enabled the robot to encourage participants to take more turns during group human-robot interactions than one of the gaze heuristic behaviors from prior work. Also, the imitation learning policy led to more active participation from human participants than another prior heuristic behavior.**

***Index Terms*—social robotics, nonverbal signals, learning**

I. Introduction

Because groups are an essential element of everyday human life, there has been a growing interest in studying how robots can interact with and in human groups [1]–[4]. Interestingly, recent work has shown that robots can assist in positive ways during group social processes [5]–[9]. However, most of these prior efforts rely on hand-crafted behavior policies based on expectations from psychology. Because group interactions are very complex, hand-crafted behaviors aiming to assist in these interactions may not always lead to the effects that were originally intended [7]–[9].

To overcome the difficulties posed by hand-crafting robot behavior, we explore the use of machine learning for robot autonomy in Human-Robot Interaction (HRI) and study whether learned behaviors can shape participation in groups. One approach to learn robot behavior is to use online learning, such as online reinforcement learning [10], but this can be risky. Random exploration or prediction errors might expose users to actions that compromise interactions. Therefore, we instead focus on learning behaviors from pre-recorded interactions.

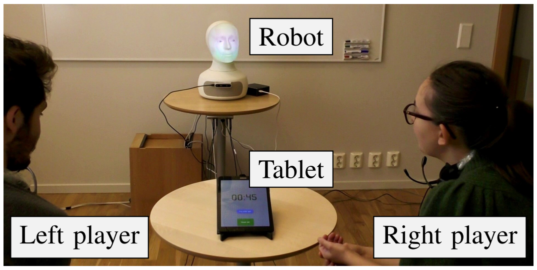

Fig. 1: The interaction scenario in which we explored learning robot gaze policies to balance human participation.

We used data collected in our previous work [11] to train non-verbal robot behavior policies to balance participation in a group HRI setting. In particular, our prior work [11] investigated how a robot's gaze could influence the behavior of two humans in a language-focused game. The humans had diverging language skill levels, which could pose challenges for collaboration and engagement in the learning environment. Nonetheless, the results showed that adaptive robot gaze could balance human participation in the game.

In this paper, we describe our efforts to learn robot gaze behaviors in groups using offline learning techniques. One approach to learning a robot policy for balancing participation is to use behavior cloning, a type of imitation learning that reduces the problem of training a policy to supervised learning [12]. In the case of the data from the language-focused game [11], one of the hand-crafted robot gaze behaviors was characterized by looking at the speaker and the listener; the other one looked at the speaker or performed gaze aversion. While the former gaze behavior was found to lead to more balanced participation in [11], gaze aversion can also help comfort a speaker [13]. Thus, we explored using imitation learning to combine strategies from both hand-crafted behaviors.

The two hand-crafted gaze policies present in the dataset from [11] might contradict each other at times, posing challenges for imitation learning. Therefore, we also explored learning a policy for balancing participation using batch reinforcement learning (RL), also known as offline RL [14]. The goal of the RL algorithm was to create a behavior policy that combined the benefits of the distinct gaze behaviors

This work was partially funded by the Swedish Research Council (no. 2017-05189), the Swedish Foundation for Strategic Research (FFL18-0199), the Jacobs Foundation (no. 2017 1261 06) and the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH.

studied in [11] while balancing participation. In comparison to imitation learning, this approach required defining a suitable reward function for training the robot policy.

We compare the proposed learning approaches in two ways. First, we use methods for policy evaluation proposed by the batch RL community. Second, we evaluate the different learned behavior policies in a between-subject study ($N = 50$) whilst keeping the interaction scenario similar to that of [11]. Fig. 1 illustrates this scenario.

In sum, our main contributions are: (1) demonstrating the use of imitation learning and batch RL for learning robot gaze policies to balance human participation in group HRI comparably to carefully hand-crafted heuristics; (2) formulating a novel objective for the RL setup that incentivizes human turn-taking during conversational interactions; and (3) extensively evaluating both learning approaches, including comparing them in a user study with respect to baseline heuristic policies from our previous work [11]. To the best of our knowledge, no prior work has studied the potential of machine learning for learning gaze policies that attempt to balance human participation in HRI.

II. Related Work

Recently, group HRI has emerged as an important field within HRI that aims to understand how robots can best interact with users in multi-party settings [1]. Also, there has been significant research in HRI on non-verbal communication, including robot gaze [15]. Due to limited space, the next sections focus on describing similar prior work to our efforts.

A. Balancing Human Participation in Group Interactions

In HRI, communicative behaviors have been shown to affect group interaction dynamics or human perception of groups [7], [8], [11], [16]–[18]. Related to our work, Tennent et al. [17] proposed a microphone-shaped robot that could balance engagement in a human group by turning towards users based on a hand-crafted policy. In terms of gaze communication, Mutlu and colleagues [18] showed that robot gaze behaviors (inspired by human gaze) can help establish conversational roles and, in turn, influence human participation behaviors. Additionally, we found in previous work that adaptive gaze behavior for a robot could aid in balancing participation in group HRI [11].

While hand-crafted robot behavior is the norm today in group HRI, there is no guarantee that these behaviors are indeed optimal for robots. Suboptimality may have multiple causes. First, group HRI is complex. For example, a robot's behavior designed to resolve human conflict led to more intense perception of the conflict in [8]. Second, in relation to creating robot gaze behaviors, much prior work in HRI is inspired by human psychology but there are gaps in this literature that can pose challenges. For instance, to the best of our knowledge, it is not well understood how human gaze can achieve a balance in participation. Further, robot and human gaze may not lead to the same responses by humans [15]. Due to the above challenges, this work explores how to leverage machine learning to generate robot behavior for group HRI.

B. Imitation Learning in HRI

Imitation learning (often known as learning from demonstration in robotics [19], [20]) has been used in the past to learn robot policies for a variety of human-robot interaction scenarios. In particular, it is common to learn robot behaviors from expert human demonstrations, such as in kinesthetic teaching of manipulation skills [21], [22]. Closer to our work, imitation learning was used by Jain et al. [23] to predict nonverbal behaviors in a conversation, including back-channelling. In contrast to these prior efforts, our aim is to learn robot behaviors using data from pre-recorded group human-robot interactions. That is, instead of having a human demonstrate behavior, we aim to learn a robot policy from the hand-crafted behaviors that were deployed in [11].

C. Reinforcement Learning in HRI

There is significant research on RL applied to HRI. For example, prior work explored how humans can teach new manipulation skills to a robot through reinforcement signals [24]. Also, RL has been used to personalize robot behavior to an interaction partner. For instance, Mitsunaga et al. [25] explored adaptive behavior to increase personal comfort based on human body signals. RL can also be used to adapt a robot's empathy [26], humor [27] and language [28] to comfort, provide entertainment, and improve learning outcomes [29].

We focus on learning robot behaviors for situated multiparty interactions. In this sense, our work is closer to that of Qureshi et al. [10], who used online RL to learn a policy for a humanoid robot to interact with bypassing strangers. Different to [10], we opted for batch RL due to the risk of random action exploration and errors that could compromise interactions with users. In this regard, [30], [31] inspired our work. These prior efforts used human interaction data and batch RL to learn non-verbal behaviors that aim to increase engagement in HRI; however, learned policies have not been deployed on a robot to the best of our knowledge.

Prior literature has also shown how to learn appropriate gaze behaviors when interacting in groups [32], [33]. We go one step beyond this line of work by aiming to learn gaze behaviors that positively shape group interaction.

III. Problem Formulation

We pose the problem of shaping participation through gaze in group HRI as a sequential decision-making problem. At any time-step t, the robot's environment is captured as a state variable $\mathbf{s}_t$. The robot can choose an action a_t that allows it to gaze towards a human, perform gaze aversion or do nothing. Our goal is to learn a gaze policy $\pi : \mathbf{s}_t \mapsto a_t$ that enables the robot to balance participation between human group members. The next sections describe the group interaction context considered in this work, our state representation, the action space, and the data used for learning and evaluating policies offline.

A. Interaction Context: A Language Game

We focus on learning gaze policies for a robot that interacts with two humans while playing a variant of the *With Other*

Words game [11]. In each round of the game, two human players describe a word via hints and the robot has to guess which word it is in a limited amount of time. The robot's role is crucial because guessing the word is a core element of the game; however, it does not require the robot to speak more than the words it guesses. Thus, the conversational interaction evolves among the two human players who have to coordinate between themselves to effectively describe words via hints. As shown in Fig. 1, the words that humans had to describe were displayed on a game tablet. The game involved a minimum of 20 words and it typically lasted 15-20 minutes.

Given the robot's goal of shaping participation, an important aspect of the interaction is *participation unevenness*. Following [17], we define unevenness in the whole interaction as:

$$\text{uneven} = \sum_{i\in[1,2]} |\text{sp}^i - \overline{\text{sp}}| \tag{1}$$

with sp^i representing the amount of time that participant i has spoken over the total amount of speech of the two human players. The term $\overline{\text{sp}}$ in eq. (1) corresponds to the mean of the relative speech time of the two players. That is, $\overline{\text{sp}} = \frac{1}{2}\sum_{i\in[1,2]} \text{sp}^i$. In this work, we use the unevenness measure to evaluate the impact of robot gaze during interactions and to learn gaze policies for the language game.

B. Interaction State

We define the state $\mathbf{s}_t$ for sequential decision making as a 77-element vector that encodes the state of human participants (36 features × 2), the state of the robot (3 features), and high-level interaction information (2 features). All of these features are collected at 2 Hz. In particular, for each human participant, the features describing their state correspond to:

Speech features: We consider speech features in our state representation because they have been useful for predicting the placement of nonverbal behaviors [23], [31]. In particular, our state includes 13-dimensional mel-frequency cepstrum coefficients (MFCC) and 4-dimensional prosody features extracted from individual audio signals. The MFCC features are computed every 25ms with a sliding hamming window of 40ms. In addition, we compute speech intensity through yin-energy and pitch through the fundamental frequency as well as the first derivative of these features. Statistical quantities are applied to feature vectors over time to describe speech over the past second. Specifically, we compute the mean and standard deviation of each feature, resulting in a 34-item feature vector.

Participation balance feature: We consider participation balance from the viewpoint of the participant. For example, let this participant be i. Then, the balance is computed in the spirit of eq. (1) but with respect to a time window $[t-w,t]$:

$$\text{uneven}^i_{[t-w,t]} = sp^i_{[t-w,t]} - \overline{sp}_{[t-w,t]} \tag{2}$$

Specifically, we use $w = 3$ minutes in this work.

Talking feature: One additional feature of the state encodes if the participant is currently talking. This feature is a binary variable (1 if the participant currently holds the floor).

The other features in $\mathbf{s}_t$ correspond to:

Robot state features: A one-hot vector of length 3 describes the current gaze target of the robot. The target can be the speaker, listener, or neither (looking away) for gaze aversion.

High-level interaction features: We consider the time since the last robot state change and the frequency of robot actions taken within one turn of a human group member. The former feature increases with time and resets to 0 every time the robot state changes. The latter provides the robot with a notion of the history of taken actions and is meant to discourage overacting.

C. Robot Gaze Actions

The actions $a_t \in A$ are discrete directions that the robot can gaze towards using its head and eye movement. More specifically, the action space A comprises four actions: *Look at speaker*, *Look at listener*, *Perform gaze aversion*, and *Do nothing*. The speaker is determined based on audio features. Gaze aversion can only be performed on the current speaker and is realized by choosing a fixed gaze target left/right above the head of the speaker. The *Do nothing* action does not change the robot's state. Note that the robot takes actions continuously, i.e., at 2Hz.

D. Interaction Data

We use data from [11] to learn gaze policies. The data comprised interactions of 26 groups in which two participants played With Other Words with a Furhat robot. Eleven groups experienced the speaker-listener condition (*SL*, originally experimental condition) from [11], in which the robot performed the actions *Look at speaker*, *Look at listener* and *Do nothing*. The other 15 groups experienced the gaze aversion condition (*GA*, originally control condition), where the robot performed the actions *Look at speaker*, *Perform gaze aversion*, and *Do nothing*. The autonomous, heuristic behaviors determined which gaze action should be executed, and for how long. The heuristic used in the SL condition used information about participation to adapt to the group. In particular, the higher the imbalance in the group, the more the robot looked at the person that had spoken the least. In the GA condition, approximately 25% of the time was spent performing gaze aversion.

A total of 2742 conversational turns are demonstrated in the dataset (1664 in GA, 1078 in SL). The recorded data contains audio streams from individual close-talk microphones, the robot's state, individual amounts of speech, and unevenness measures. While our work focuses on using this dataset to train robot gaze behaviors, one could imagine using other sources of data in the future given the flexibility of learning algorithms. This possibility is further discussed in Section VII-G.

IV. Estimating a Robot Gaze Policy with Imitation Learning

We explore learning a gaze policy for a robot using imitation learning (IL) and the group interaction data from [11]. As explained in Section III-D, the data was collected when the robot showed two distinct gaze behaviors. In one case, the robot switched between gazing at the speaker and the listener. This behavior was shown to reduce participation unevenness among human interactants. In the other case, the robot looked only towards the speaker or looked away, performing gaze aversion. Because humans often use gaze aversion to avoid staring at other people and comfort speakers during conversations [13], we used the data from both cases to learn a gaze policy. We hoped that the learned policy would help balance participation while seeming natural to users.

We learned the policy using behavioral cloning [12]. That is, we used supervised learning to map observed states $\mathbf{s}_t$ to actions a_t given paired input-target data from [11]. In particular, we investigated how well three types of policy models worked for rendering a suitable robot gaze behavior: a decision tree (DT) classifier, a linear classifier trained via stochastic gradient descent (SGD) [34], and a k-nearest neighbors (KNN) classifier [35]. The decision tree classifier was trained using the CART algorithm [36] with the Gini impurity criteria. We considered a range of parameters for maximum tree length, minimum number of samples per node and minimum impurity decrease to split a node. For the SGD-based models, we considered different loss functions (e.g., hinge loss) and regularization norms. Finally, for the KNN classifier, we considered different algorithms (e.g., Ball Tree [37], KD Tree [38]), leaf sizes and distance metrics (L1, L2). We used grid search to find suitable parameters for these models using the Scikit-learn library.[1]

V. Estimating a Robot Gaze Policy with Batch Reinforcement Learning

Behavioral cloning as explored in Section IV generally assumes that an expert provides consistent examples for learning, i.e., the "ground truth". However, our data was obtained from interactions in which the robot executed two gaze behaviors. We suspected that this could pose challenges for learning a suitable gaze policy. Thus, we decided to also explore RL for this problem. A key difference between IL and RL is that RL can potentially learn a better policy from mistakes given a suitable reward function, rather than simply *copying* behavior.

To estimate a suitable robot gaze policy via batch RL, we first defined a relevant horizon H for the problem as well as a reward function r_t. To define the horizon, we closely studied the problem of achieving balanced participation. First, we observed that an essential aspect of this problem is turn-taking. Intuitively, we wanted the robot to subtly incentivize switches in the conversational floor so that all human interactants get a chance to express themselves. Second, we expected learning of a policy to be difficult if the horizon equaled the full length of a group interaction (i.e., the full length of a With Other Words game). This expectation was based on prior work that has shown the difficulty of RL problems scales in terms of sample complexity as the horizon increases [39]. Given the limited amount of interaction data that is generally available for group HRI, we decided to frame our RL problem as estimating a suitable robot gaze policy for one conversational turn. In other words, an episode in our RL setup starts when a person takes the turn to speak and ends when that person releases the floor.

To explain the rationale for our reward function, we make two observations. First, if possible, the current turn should improve the balance of the conversation. Second, many short turns without meaningful spoken contributions are undesired. Therefore, we propose a reward r_t that describes the quality of the interaction in regards to human participation balance subject to the length of the current turn. More formally, let t be the time when the current turn ends, i.e., when there is a change in the conversational floor. Also, let $t - l$ be the time when the prior turn ended, where l corresponds to the length of the current turn. Then, the proposed reward is:

$$r_t = (\underbrace{\text{uneven}_{[t-l-w,t-l]}}_{\substack{\text{unevenness at the}\\ \text{end of the prior turn}}} - \underbrace{\text{uneven}_{[t-w,t]}}_{\substack{\text{unevenness at the}\\ \text{end of the current turn}}}) * l \qquad (3)$$

The reward is given only at the end of a turn due to the importance of changes in the conversational floor. For all other time-steps, the reward is zero.

We learn a policy using the Double Deep Q-learning algorithm [40], which approximates the optimal Q-value function: $Q^*(\mathbf{s}, a) = \mathbb{E}_{\pi^*}[R_t \,|\, \mathbf{s}_t = \mathbf{s}, a_t = a]$. This function corresponds to the expected return starting from a state $\mathbf{s}$, taking the action a, and thereafter acting optimally [41]. Prior work has shown that estimating the Q function without the possibility of exploration can cause an extrapolation error resulting in an unrealistic estimation of the Q value for unseen state-action pairs [42]. Thus, the implementation we use constraints the learned Q function as in Batch-Constrained Deep Q-learning [42], so that while learning it can only consider future actions for which the state-action pair is in the data.

The neural network architecture that we used to predict the Q function was composed of three fully connected (FC) layers with 256, 512 and 4 units, respectively. The first two FC layers were followed by ReLU activations. Our implementation leveraged the Coach RL library to train this model.[2]

VI. Training Gaze Policies and Offline Evaluation

We evaluate the proposed IL and RL approaches for learning robot gaze policies quantitatively using a variety of metrics.

A. Train, Validation, Test Datasets

We trained and evaluated models using the interaction data described in Section III-D. Because there was no specific order for individual participant data to be included in the state feature vector, we augmented the dataset by creating states for the two different possible placements of individual participant data. For RL, this resulted in a total of 5484 episodes where one episode comprised one turn of the original interaction.

We split the episode data into a training set (approximately 48%), a validation set (32%) for choosing hyper-parameters, and a test set (20%) for analyzing chosen policies on unseen interactions prior to human evaluation. The same train/validation/test splits were used for both IL and RL. The features in the state vectors were normalized before training.

[1] https://scikit-learn.org/

[2] https://github.com/IntelLabs/coach

TABLE I: Metrics from test data in imitation models. Macro F1-score describes the top performing 10% of models for each method. The decision tree model was selected for deployment.

Model Type	Macro F1 score ($M \pm SD$)	WIS
DT	0.359 ± 0.005	0.124
SGD	0.335 ± 0.003	-0.002
KNN	0.278	-0.013

B. Evaluation Metrics

Because the selection of the random seeds can strongly affect the training process and is important for the reproduction of scientific results, we evaluated different behavior policies obtained with different seeds. An initial evaluation was performed during training that used the macro F1 metric (for IL) and the mean of the Temporal Difference error (for RL). Once models were trained, we evaluated them using additional off-policy evaluation metrics from the RL community. For completeness, we now provide high-level descriptions of all of these metrics. We encourage interested readers to consult the original works for more details.

Macro F1 Score: The macro F1 score is built as an average of the F1 score for each of the classes. The F1 score thereby computes the harmonic mean between precision and recall.

Mean of Temporal Difference (TD) Error: The TD error reflects the difference between the Q value for a state-action pair estimated through the Bellman equation and the prior estimate of the Q value [41, Ch. 6.1]. With the mean TD error, we refer to the mean squared error that is built over the TD error for all state-action pairs.

Weighted Importance Sampling (WIS): This metric is used to estimate the expected returns under the learned policy given samples of the behavior policy (used to generate the dataset). WIS uses a notion of relative probability between the two policies to reweigh the reward obtained in the episodes present in the dataset [41, Ch. 5.5].

Sequential Doubly Robust (SDR): This metric is an application of the doubly robust estimator to sequential decision-making problems [43]. This means that it uses two techniques to estimate the average value of the learned policy: one based on the learned Q-function and one based on the relative probabilities (see WIS), the reward and the prior SDR estimation.

Note that we selected WIS for our evaluation because it allowed us to estimate the quality of a policy without an estimate of the Q function and therefore also can be used to estimate the quality of an IL policy. Further, we considered SDR because it is unbiased and has lower variance than WIS, helping judge the quality of RL policies.

Because randomness was involved in the gaze behavior for both conditions from [11], the action probabilities needed to calculate the relative probabilities for WIS and SDR could not be extracted directly from the data. Instead, we calculated action probabilities as needed for the RL metrics above using approximate nearest neighbors, as suggested in [44].

C. Training of Imitation Learning Policies

The decision tree and the linear classifier fitted through SGD were trained using 51 different seeds and using grid search to find the best hyperparameters. The top 10% of candidates for each model were initially selected according to their macro F1 score on the validation data. This score was computed not on the actions commanded to the robot – which were used for training the mapping from state to actions – but rather on the robot gaze state because it was less sparse than the actions and better reflected the robot's gaze target over time.

The final IL policy was selected out of the top 10% of candidates based on its WIS score. As shown in Table I, this corresponded to a decision tree model, which used the Gini impurity and a minimum of two samples to split a node.

It is worth noting that the F1 scores for the IL policies are low overall in Table I. This was expected because the validation dataset contains episodes from the GA and SL conditions from [11], where only 2 out of the 3 possible gaze behaviors are present. However, the IL agent could predict any of the 3 behaviors because it combined the heuristics into a single policy. Note that prior to deployment, we verified the behavior output by the chosen IL policy in the test dataset, as explained below in Section VI-E.

D. Training of Batch RL Policies

We obtained gaze policies via batch RL using an episodic memory from which the training batches were selected. The training process for the Q-value function lasted a total of 58500 steps, distributed over 150 training epochs. The initial learning rate was 0.006 and decayed exponentially every 10000 steps. The final learning rate was $1.258e^{-5}$.

Because the network representing the Q function is randomly initialized, we trained RL policies with 51 different random seeds. For each of these training events, we identified the best policy across epochs by inspecting their WIS, SDR and mean TD error values. In particular, we first identified the training epoch for each model that resulted in the highest values of WIS and SDR on the validation set. Then, we narrowed down our search for a model by choosing the top 10% of models based on their TD error in the training set. Lastly, we chose an RL model as main candidate for our user study (Section VII) by ranking models based on their WIS and SDR on our validation set. The WIS of the best performing model had a value of 0.0089. Using the WIS allowed us to directly compare the policies learned through imitation learning and batch reinforcement learning.

E. Analysis of Inferred Actions on the Test Set

To get a better understanding of the candidate RL and IL policies for our user study, we predicted their actions on the test dataset, which was unseen at training time. That is, we rolled out the policies and updated the state space according to the predicted gaze actions. Then, we evaluated for how long the robot stayed in the different states. Further, we evaluated the frequency of state changes – which we denoted as *hastiness*

TABLE II: Comparison of different gaze policies for the test dataset ($M \pm SD$ are calculated over episodes). Action columns refer to the relative time spent in the robot state that was reached after taking the respective action.

Condition	**Hastiness**	**Duration of actions (s)**	***Look at speaker***	***Look at listener***	***Perform gaze aversion***
GA	0.164 ± 0.012	3.054 ± 3.971	0.832 ± 0.010	-	0.168 ± 0.010
SL	0.134 ± 0.008	3.738 ± 2.681	0.743 ± 0.001	0.257 ± 0.001	-
RL	0.226 ± 0.031	2.219 ± 3.221	0.741 ± 0.080	0.089 ± 0.041	0.169 ± 0.074
IL	0.114 ± 0.030	4.421 ± 6.052	0.815 ± 0.039	0.113 ± 0.030	0.072 ± 0.065

– as well as the frequency of states. This analysis helped evaluate intermediate results during the development process.

Table II summarizes our findings on the test dataset considering the best RL and IL policies chosen earlier (Sections VI-D and VI-C) and the heuristic gaze behaviors from [11] as a reference. The batch RL gaze policy appeared to be hastier than the other policies. That is, it changed gaze targets more frequently. The imitation learning policy appeared slightly less hasty than the heuristics, although it gazed towards the listener for less time than the behavior from the SL condition from [11]. We considered the results to be reasonable, allowing for the deployment of the final behavior policies on the robot.

VII. User Study

We conducted a between-subject study with two conditions to compare the IL and the RL gaze policies chosen via our offline evaluation (Section VI). Further, this study served to compare the policies with the heuristic gaze behaviors originally proposed in [11] to balance human participation in group HRI. In general, our study protocol followed the protocol from [11], including the recruitment of native Swedish speakers and Swedish language learners to play the With Other Words game with a Furhat robot (as depicted in Fig. 1).

A. Participants

For the present study, 50 participants were recruited at a university campus and surrounding areas through flyers, posters, word of mouth and social media platforms. Participants' ages ranged from 18 to 67 years ($M = 28.02$, $SD = 10.53$), and 27 participants identified as female, 22 as male and 1 would rather not say. Thirty-two participants reported that they had never interacted with a robot before, and 8 indicated interacting with robots regularly. The participants were grouped into dyads for our study. One pair of participants indicated being connected via social media; all had never met before.

We analyzed data for our 50 participants along with the data of 72 participants who took part in the study from [11]. Considering all 122 participants, there were 61 dyads who interacted with the Furhat robot. However, of the 25 groups formed by the 50 participants newly recruited for this study, 1 group had to be excluded because the participants misinterpreted the instructions. In [11], 9 of their 36 groups also had to be excluded. This meant that we had a total of 102 valid participants, corresponding to 51 dyads. Table III presents demographic details for these 102 participants.

TABLE III: The demographics of the participants (NS: participants who rather did not share their gender identity). The language level information only covers the language learners. The remaining participants were native Swedish speakers.

Condition	**Age**	**Gender**			**Swedish language level**			
		F	**M**	**NS**	**A2**	**B1**	**B2**	**C1**
RL	27.2 ± 11.9	14	9	1	3	5	2	2
IL	28.6 ± 9.9	12	12	0	2	6	2	2
GA	32.0 ± 10.7	12	18	0	3	4	6	2
SL	31.8 ± 11.5	12	12	0	1	5	4	2

B. Study Design & Hypotheses

We conducted the study with a between-subject design, as in [11]. We collected data for two conditions: IL and RL, using the policies obtained from our offline evaluation (Section VI). Further, we considered the data collected in [11] as two additional conditions: SL and GA. The SL condition used an adaptive gaze strategy to look between the speaker and the listener. In the GA condition, the robot only focused on the speaker and performed gaze aversion to keep the dynamics of the gaze comparable to the SL condition. As in the IL and RL conditions, the robot operated fully autonomously in the SL and GA conditions. However, its behavior in the latter two conditions was based on heuristics rather than learned. Considering all four conditions, we hypothesized:

H1 *The IL and the RL policies will result in lower participation unevenness than the GA and SL policies.*

We expected the above result because both the IL and RL policies tried to leverage the data from the other conditions to achieve balanced participation.

H2 *The RL policy will result in lower unevenness than the IL policy.*

This second hypothesis was motivated by the fact that the RL policy was trained on a reward function that was directly tied to the notion of participation unevenness whereas this notion was implicit when training the IL policy.

H3 *The RL policy will result in more turns taken by the participants compared to the other three policies.*

This last hypothesis was driven by our RL setup. The batch RL agent received a reward at the end of a turn and, thus, it could potentially benefit from trying to switch speakers early. Also, our offline evaluation on a subset of the data from [11] suggested that RL could lead to higher hastiness (Table II).

As part of the study, we also investigated how participants perceived the robot and reacted to it in the different conditions.

C. Study Protocol

After giving written consent, the participants were asked to complete a demographics questionnaire. We assigned participant dyads (one native Swedish speaker, one language learner) through blocked randomization to one of the IL or RL

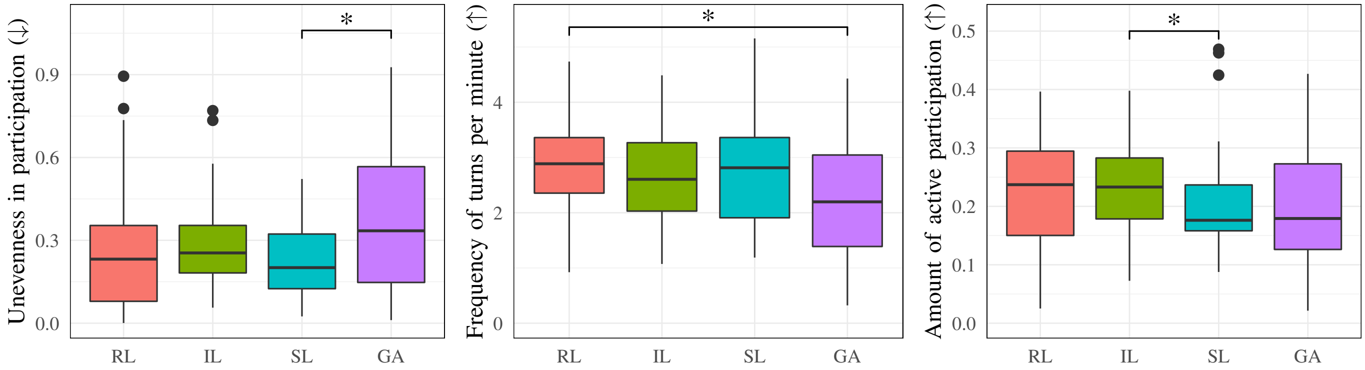

Fig. 2: Results from the analysis of unevenness of participation, number of turns, and active amount of participation. ↑ indicates that high values on this measure are desirable, and ↓ indicates that low values reflect a successful interaction. Horizontal bars represent the median and the box represents the inter-quartile range. The * symbol indicates $p < 0.05$.

conditions. This was important to ensure a similar distribution of language levels among language learners. Participants were then invited to a conference room to meet the Furhat robot.

During the study, the robot introduced itself, the participants were invited to do the same, and the robot then explained the game rules. Afterwards, the participants played With Other Words with Furhat for 15-20 minutes and, finally, completed a post-game questionnaire away from the robot. This included answering questions about participants' familiarity and their perceptions of the robot. Before debriefing about our research interests in robot gaze, the participants were compensated for their participation with a voucher (value ∼11.5 USD).

D. *Measures*

1) In-Game Measures: During With Other Words, the participants were asked to describe words with increasing levels of difficulty. Thus, each of the following measures were collected on each level separately and normalized to the time spent on the respective level.

Active Participation: We collected the amount of active participation through voice activity detection for each participant using individual close-talk microphones.

Frequency of Turn-Taking: As the number of turns taken among participants is co-dependent, we measured how frequently the participants took turns on a group level. The system assumes that a turn was taken when the binary voice activation signal indicated a change in speaker that lasted longer than a backchannel threshold of 1 second.

Unevenness of Participation: We used eq. (1) to calculate the unevenness of participation at each difficulty level. Low values indicate an even speech distribution and balanced participation.

As in [11], the robot's voice could sometimes be heard in the participants' close-talk microphones. But, due to a slight difference in participants' positions in our setup (Fig. 1) and that of [11], we worried about additional false detection of speech in the RL and IL conditions. Therefore, we applied a correction to our audio processing pipeline that adjusted for this issue. For fairness, this correction was applied equally to data from the IL and RL as well as SL and GA conditions.

2) Subjective measures: We collected characteristics of our participants and perceptions of the robot in pre- and post-game questionnaires. These measures included participant **Extroversion** and **Agreeableness** from the Big Five Inventory [45], [46] as we expected these traits to influence human behavior. Also, we measured **Willingness to Communicate** [47] and participants' Swedish **proficiency** [48].

We measured impressions of the robot with the **Warmth** and **Discomfort** scales from the Robotic Social Attributes Scale (RoSAS) [49]. Lastly, because we wanted to understand if the learned gaze behaviors are perceived as unpleasant, we asked the participants to rate on a 5-point scale the degree to which they found the robot **intimidating**, if they thought it was **ignoring** them, or whether it was **staring** at them.

E. *Results*

1) Participation Unevenness: We performed a one-way ANCOVA to examine the effects of *condition* on the *unevenness of participation*, after controlling for the *proficiency of the language learner* because we expected it to influence participation behavior. The analysis yielded a main effect of condition for the unevenness of participation measure, $F(3, 146) = 2.98, p = 0.034$. A post-hoc pairwise comparison with Tukey HSD correction showed that the GA and SL condition were significantly different, $p = 0.022$, as reported in our prior work [11] (SL ($M \pm SD$): 0.23 ± 0.13, GA: 0.37 ± 0.25). Fig. 2 (left) shows these results.

Additionally, the language proficiency covariate was significantly related to the *unevenness in participation*, $F(3, 146) = 4.133$, $p = 0.007$. This indicated that the imbalance in skill induced by different language levels influenced participation.

2) Frequency of turns taken: We analyzed the effect of *condition* on the *number of turns taken* through one-way ANCOVA while controlling for the *proficiency of the language learner*. The analysis showed a main effect of condition on the number of turns taken, $F(3, 146) = 3.036$, $p = 0.031$. A post-hoc pairwise comparison test with Tukey HSD correction indicated that the participants in the RL condition took more turns than those in the GA condition, $p = 0.024$

(RL ($M \pm SD$): 2.86 ± 0.86, GA: 2.27 ± 1.05). The covariate language proficiency showed a trend towards significance for the number of turns taken, $F(3, 146) = 2.599$, $p = 0.054$. More details are given in the middle plot of Fig. 2.

3) Amount of active participation: To understand the effect of *condition* on the *amount of active participation*, we performed a one-way ANCOVA after controlling for *willingness to communicate*, *language proficiency* (4 learner levels + native speaker), and the personality traits of *agreeableness* and *extroversion* because we expected those participant characteristics to influence participation behavior. The analysis showed a main effect of condition on the active amount of participation, $F(3, 304) = 3.655$, $p = 0.013$. A post-hoc pairwise comparison test with Tukey HSD correction indicated that the participants in the IL condition participated more actively (e.g., speech, laughter, back-channeling) than participants in the SL condition, $p = 0.007$ (IL ($M \pm SD$): 0.234 ± 0.086, SL: 0.217 ± 0.088). The covariates extroversion ($F(1, 304) = 9.643$, $p = 0.012$) and language proficiency ($F(4, 304) = 32.190$, $p < 2.2e - 16$) had a significant effect on the active amount of participation. The right plot in Fig. 2 shows active participation by condition.

4) Robot perception: Finally, we conducted a one-way ANOVA to study the effect of robot gaze behavior on perceptions of the robot (Warmth, Discomfort, Ignoring, Intimidating, and Staring). The analysis showed no effect of condition on the perceptual measures. Ratings were generally positive.

We further analyzed the participant's comments about the robot in the post-experiment questionnaire. Note that this qualitative data was only available for the RL and IL conditions. We found that 50% of the participants did not notice anything unusual about the robot's behavior. Four participants (8%) noted negative gaze behaviors (3 RL; 1 IL), mostly reporting hasty movements. Other comments positively and negatively assessed the guessing (6 responses) or the hard-coded behaviors, such as back-channelling (11). When asked specifically to describe the robot's gaze behavior, 25% characterized the robot's gaze as friendly or kind, 15% described it as attentive (6 RL, 1 IL), 19% as sharing attention equally (3 RL, 6 IL), 12.5% as looking at the speaker (1 RL, 5 IL), and 6% noted rapid and hasty movements (3 RL, 0 IL). The remaining comments concerned the gaze intensity (3) or were general (8).

F. Discussion of the Results

One of our goals was to learn gaze behaviors that can balance participation. We did not find that the learned behaviors improved the balance in the interactions with respect to prior work [11] (**H1**) nor that RL led to lower unevenness than IL (**H2**). However, there was also no indication that the learned gaze policies would decrease the balance and quality of interactions. Also, we found partial support for our hypothesis (**H3**) that the gaze behaviors trained through batch RL and deployed on the robot could increase the number of turns for human participants in the game. In particular, the RL gaze policy led to more turns than the GA behavior from [11]. We attribute this increase in turns to our RL problem formulation and our novel reward. Because turn-taking is important for group collective intelligence [50], we argue that the RL behavior shows promise towards improving the quality of group HRI. Interestingly, we also found that IL led to more active participation than the SL condition, showing potential for this method as well. Overall, these result indicate that learning of gaze policies was successful in the sense that the learned behaviors did not compromise the interaction and helped with turn taking or verbal participation.

G. Limitations

First, we studied interactions with an autonomous robot, and its signal processing software sometimes missed or detected false human speech. While we corrected for this issue in our study measures, this could have affected the real-time behavior of the robot in our human evaluation. Second, our study recruitment procedure only allowed us to apply randomized blocking to one factor (language learning). The participants' gender distribution differed between our study conditions and might have influenced our findings. Third, we learned and tested robot gaze policies in a controlled interaction scenario. It would be interesting to adapt the proposed methods to in-the-wild human-robot interactions, where more than two people interact with the robot. Lastly, it would be interesting to investigate if the proposed approaches can effectively learn from other types of data than the heuristic gaze behaviors from [11], e.g., data collected in other cultures or other group tasks. Given the popularity of using human data for learning robot behaviors in HRI, it would also be interesting to understand if such data can serve for balancing participation in group HRI via gaze. Assuming this data was available, future work should consider (a) possible differences in human reaction to robot vs. human gaze [15] and (b) the mapping of human gaze and head motion to robot motions, which is nontrivial due to physical characteristics of the hardware [51].

VIII. Conclusion

Our work demonstrated that imitation learning, in the form of behavorial cloning, as well as batch RL were suitable to combine robot gaze behaviors that aim to balance participation in group HRI. For RL, we proposed a novel reward and chose a horizon that was intended to encourage balancing participation. An extensive offline evaluation allowed us to evaluate the learned policies quantitatively to ensure that learning resulted in reasonable behaviors before deployment. Our user study showed that the learned policies shaped participation behavior. While the participation balance achieved was similar, the learned behaviors influenced the amount of turn-taking (RL) and active participation (IL) when comparing to the hand-crafted heuristics that were used to generate the human-robot interaction data. This showed promise in learning gaze behaviors for group HRI. In the future, we foresee using similar learning methods to those proposed in this paper to create suitable robot gaze policies to shape other group interactions and to explore learning other non-verbal behaviors in an offline fashion from group HRI data.

References

[1] S. Sebo, B. Stoll, B. Scassellati, and M. F. Jung, "Robots in Groups and Teams: A Literature Review," *Proc. ACM Hum.-Comput*, vol. 4, no. October, p. 37, 2020. [Online]. Available: https://doi.org/10.1145/3415247

[2] M. R. Fraune, S. Sabanovic, and E. R. Smith, "Teammates first: Favoring ingroup robots over outgroup humans," *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, vol. 2017-Janua, no. August, pp. 1432–1437, 2017.

[3] F. Correia, S. Mascarenhas, R. Prada, F. S. Melo, and A. Paiva, "Group-based Emotions in Teams of Humans and Robots," *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, no. February, pp. 261–269, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3171221.3171252

[4] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards Robot Autonomy in Group Conversations," *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pp. 42–52, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2909824.3020207

[5] S. Gillet, W. van den Bos, and I. Leite, "A social robot mediator to foster collaboration and inclusion among children," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[6] S. Strohkorb Sebo, L. L. Dong, N. Chang, and B. Scassellati, "Strategies for the Inclusion of Human Members within Human-Robot Teams," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 3 2020, pp. 309–317. [Online]. Available: https://dl.acm.org/doi/10.1145/3319502.3374808

[7] S. Strohkorb, E. Fukuto, N. Warren, C. Taylor, B. Berry, and B. Scassellati, "Improving human-human collaboration between children with a social robot," *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pp. 551–556, 2016.

[8] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using Robots to Moderate Team Conflict: The Case of Repairing Violations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. Portland, Oregon, USA: Association for Computing Machinery, 2015, p. 229–236. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2701973.2702094

[9] E. S. Short, K. Swift-Spong, H. Shim, K. M. Wisniewski, D. K. Zak, S. Wu, E. Zelinski, and M. J. Mataric, "Understanding social interactions with socially assistive robotics in intergenerational family groups," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 236–241, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8172308/

[10] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network," *arXiv*, pp. 1639–1645, 2017.

[11] S. Gillet, R. Cumbal, A. Pereira, J. Lopes, O. Engwall, and I. Leite, "Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 3 2021. [Online]. Available: https://doi.org/10.1145/3434073.3444670

[12] D. Pomerleau, "An autonomous land vehicle in a neural network," *Advances in Neural Information Processing Systems; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA*, 1998.

[13] A. Abele, "Functions of gaze in social interaction: Communication and monitoring," *Journal of Nonverbal Behavior*, vol. 10, no. 2, pp. 83–101, 1986.

[14] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement learning*. Springer, 2012, pp. 45–73.

[15] H. Admoni and B. Scassellati, "Social Eye Gaze in Human-Robot Interaction: A Review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, p. 25, 3 2017. [Online]. Available: http://dl.acm.org/citation.cfm?id=3109975

[16] S. Strohkorb Sebo, M. Traeger, M. F. Jung, and B. Scassellati, "The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams," *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, no. February, pp. 178–186, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3171221.3171275

[17] H. Tennent, S. Shen, and M. Jung, "Micbot: A Peripheral Robotic Object to Shape Conversational Dynamics and Team Performance," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2019-March, pp. 133–142, 2019.

[18] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, pp. 1–33, 2012.

[19] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[20] A. Billard and D. Grollman, "Robot learning by demonstration," *Scholarpedia*, vol. 8, no. 12, p. 3824, 2013.

[21] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 391–398.

[22] C. L. Mueller and B. Hayes, "Safe and robust robot learning from demonstration through conceptual constraints," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 588–590.

[23] V. Jain, M. Leekha, R. R. Shah, and J. Shukla, "Exploring semi-supervised learning for predicting listener backchannels," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–12.

[24] A. L. Thomaz and C. Breazeal, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," *Proceedings of the National Conference on Artificial Intelligence*, vol. 1, pp. 1000–1005, 2006.

[25] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Robot Behavior Adaptation for Human-Robot Interaction based on Policy Gradient Reinforcement Learning," *Journal of the Robotics Society of Japan*, vol. 24, no. 7, pp. 820–829, 2006.

[26] I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, and A. Paiva, "Modelling Empathy in Social Robotic Companions," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7138 LNCS, no. July, pp. 135–147. [Online]. Available: http://link.springer.com/10.1007/978-3-642-28509-7_14

[27] K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. André, "How to shape the humor of a robot - Social behavior adaptation based on reinforcement learning," *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pp. 154–162, 2018.

[28] H. Ritschel, T. Baur, and E. Andre, "Adapting a Robot's linguistic style based on socially-Aware reinforcement learning," *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, vol. 2017-Janua, pp. 378–384, 2017.

[29] G. Gordon, S. Spaulding, J. Korywestlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, no. 2011, pp. 3951–3957, 2016.

[30] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Batch recurrent q-learning for backchannel generation towards engaging agents," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[31] N. Hussain, E. Erzin, T. Metin Sezgin, and Y. Yemez, "Speech driven backchannel generation using deep Q-network for enhancing engagement in human-robot interaction," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 4445–4449, 2019.

[32] M. Vazquez, A. Steinfeld, and S. E. Hudson, "Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach," *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pp. 36–43, 2016.

[33] S. Lathuilière, B. Massé, P. Mesejo, and R. Horaud, "Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction," *Pattern Recognition Letters*, vol. 118, pp. 61–71, 2019.

[34] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, Y. Lechevallier and G. Saporta, Eds. Paris, France: Springer, August 2010, pp. 177–187. [Online]. Available: http://leon.bottou.org/papers/bottou-2010

[35] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879

[36] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and regression trees,” The Wadsworth Statistics/Probability Series. Belmont, California: Wadsworth International Group, a Division of Wadsworth, Inc. X, 358 p. $ 29.25; $ 18.95 (1984)., 1984.

[37] S. M. Omohundro, “Five balltree construction algorithms,” Tech. Rep., 1989.

[38] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” *Commun. ACM*, vol. 18, no. 9, p. 509–517, Sep. 1975. [Online]. Available: https://doi.org/10.1145/361002.361007

[39] R. Wang, S. S. Du, L. F. Yang, and S. M. Kakade, “Is long horizon reinforcement learning more difficult than short horizon reinforcement learning?” *arXiv preprint arXiv:2005.00527*, 2020.

[40] H. v. Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI’16. AAAI Press, 2016, p. 2094–2100.

[41] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2018. [Online]. Available: https://books.google.se/books?id=6DKPtQEACAAJ

[42] S. Fujimoto, D. Meger, and D. Precup, “Off-policy deep reinforcement learning without exploration,” in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2052–2062. [Online]. Available: https://proceedings.mlr.press/v97/fujimoto19a.html

[43] N. Jiang and L. Li, “Doubly robust off-policy value evaluation for reinforcement learning,” in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 652–661. [Online]. Available: https://proceedings.mlr.press/v48/jiang16.html

[44] A. Raghu, O. Gottesman, Y. Liu, M. Komorowski, A. Faisal, F. Doshi-Velez, and E. Brunskill, “Behaviour policy estimation in off-policy policy evaluation: Calibration matters,” *arXiv preprint arXiv:1807.01066*, 2018.

[45] O. P. John, E. M. Donahue, and R. L. Kentle, “The big five inventory—versions 4a and 54,” 1991.

[46] O. P. John, L. P. Naumann, and C. J. Soto, “Paradigm shift to the integrative big five trait taxonomy,” *Handbook of personality: Theory and research*, vol. 3, no. 2, pp. 114–158, 2008.

[47] S. Ryan, “Self and identity in l2 motivation in japan: The ideal l2 self and japanese learners of english,” *Motivation, language identity and the L2 self*, vol. 120, p. 143, 2009.

[48] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001.

[49] 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.

[50] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, “Evidence for a Collective Intelligence Factor in the Performance of Human Groups,” *Science*, vol. 330, no. 6004, pp. 686–688, 10 2010. [Online]. Available: https://www.science.org/doi/10.1126/science.1193147

[51] P. Agarwal, S. Al Moubayed, A. Alspach, J. Kim, E. J. Carter, J. F. Lehman, and K. Yamane, “Imitating human movement with teleoperated robotic head,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 630–637.