

Shielding for Socially Appropriate Robot Listening Behaviors

Sarah Gillet¹, Daniel Marta¹, Mohammed Akif¹ and Iolanda Leite¹

Abstract—A crucial part of traditional reinforcement learning (RL) is the initial exploration phase, in which trying available actions randomly is a critical element. As random behavior might be detrimental to a social interaction, this work proposes a novel paradigm for learning social robot behavior—the use of shielding to ensure socially appropriate behavior during exploration and learning. We explore how a data-driven approach for shielding could be used to generate listening behavior. In a video-based user study (N=110), we compare shielded exploration to two other exploration methods. We show that the shielded exploration is perceived as more comforting and appropriate than a straightforward random approach. Based on our findings, we discuss the potential for future work using shielded and socially guided approaches for learning idiosyncratic social robot behaviors through RL.

I. INTRODUCTION

The initial stage of training a reinforcement learning (RL) agent is often based on random exploration [1]. However, a randomly acting robot might be inappropriate, be it while trying to grasp some object in the environment or in a social interaction with a human. In the latter example, the robot might nod continuously or nod even though nothing was said by the person. These inappropriate behaviors could affect the human in the interaction and alter the interaction. A socially inappropriate, randomly acting robot cannot guarantee realistic interactions from which to learn the desired behavior. Therefore, we need to ensure that an RL robot acts socially appropriately during exploration and learning.

In this work, we propose a novel paradigm for learning social robot behavior that leverages techniques from safe RL, i.e., shielding [2], to ensure socially appropriate behaviors. Prior work has approached learning social robot behavior through, e.g., imitation learning [3] or reinforcement learning [4], [5]. To ensure appropriateness, related work trained offline or evaluated policies on unseen data before deployment [6]. However, offline approaches require exhaustive and consistent enough datasets to allow for successful training.

In this paper, we address this challenge in the context of generating socially appropriate listening behavior for a social robot. An active listening robot should perform short vocal or non-vocal backchannels [7], e.g., paraverbals (‘mm-hmm’, ‘uh-huh’, etc.) or nod the head. The *timing* of these backchannels and the *type* of backchannel, vocal or non-vocal, determine if a backchannel is appropriate. In this work,

This work was supported by the S-FACTOR project from NordForsk, the Swedish Foundation for Strategic Research (SSF FFL18-0199), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH

¹All authors are with KTH Royal Institute of Technology, Lindstedsvägen 24, 10044 Stockholm, Sweden, sgillet@kth.se

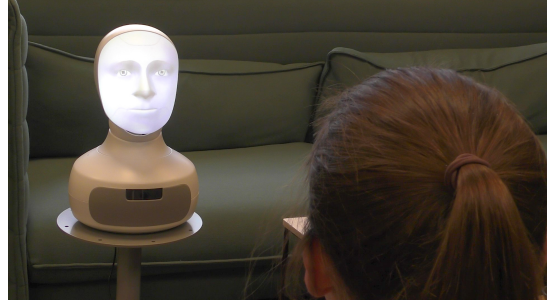


Fig. 1: Snippet from the footage used for the video study. The experimental setup implies a conversation between the robot Furhat and a person. In the shown video, the robot indicates listening through backchanneling utterances and nods.

we focus on ensuring the appropriate *timing* of backchannels as one contributing factor to socially appropriate behavior.

To address the challenge of timing for backchanneling behavior, we build a data-driven shield for our randomly exploring reinforcement learning robot combined with the concept of Backchannel Opportunity Points (BOP) [8]. The safe reinforcement learning community often uses a set of rules or guarantees to realize shielding and ensure safe behavior. However, prior work exploring the imitation of human backchanneling behavior showed that data-driven backchanneling behavior outperforms rule-based behavior [9]. In addition, to the best of our knowledge, no prior work explored a set of rules for identifying socially appropriate timing opportunities for backchannels, i.e., BOPs, which could function as a shield. Therefore, we trained a regression model on a conversational dataset and combined it with a threshold on the output to implement the shield.

In a video-based human-robot interaction study, we show that the exploring robot with a shield is perceived as a significantly better listener and as acting more appropriately than the completely random robot. Interestingly, a statistically guided randomly behaving robot was perceived as the least rude compared to the shielded and completely random exploring robot. We discuss these findings in relation to the reinforcement learning process. In summary, our contributions are as follows:

- 1) we propose a novel paradigm for generating socially appropriate listening behavior within a reinforcement learning robot using shields
- 2) we explore data-driven techniques by training a regression model and combining the output with a threshold to implement shielding
- 3) we evaluate the best shield in comparison to other exploration approaches in a video-based user study (N=110).

II. RELATED WORK & BACKGROUND

Our work builds upon literature on backchanneling and the generation of listening behavior, as well as, work that explores RL and shielding in the context of HRI. Listening behavior is characterized by backchannels (BCs) [7], which are portrayed through body language, such as head nods or smiles, and short vocalizations ('uh-huh', 'hmm'). Backchannels are important in conversations, as they can skew the perception of the interactants [10], including robots [11].

Backchanneling Opportunity Points (BOP). Backchanneling occurs in Backchanneling Opportunity Points (BOP) [8], sometimes also called backchanneling relevant spaces. Previous work indicates that, on average, there are 3.5 times more BOPs than actual backchannels [12]. In addition, recent research has shown that backchanneling behavior is idiosyncratic [13], i.e., it is peculiar to an individual, differing significantly. Therefore, it is particularly interesting to explore how robots could learn their backchanneling behavior whilst ensuring socially appropriate behaviors through shielding.

Generating listening behavior. Significant amount of work has studied the generation of listening behavior in robots. Several approaches explored rule-based methods for placing backchannels informed by, e.g., prosodic features [14], [15], or automated methods, e.g., with the help of Hidden Markov Models and prosodic [16] or multimodal features [17]. More recently, deep learning techniques have been used to generate listening behaviors, e.g., with the help of Recurrent Neural Network models, which capture the temporal dependencies of continuous signals (i.e., retain “memory” of previous inputs) [18]. Prior work explored Long-Short Term Memory (LSTM) layers with audio signals and word history [19] or made use of LSTMs and multimodal input [9], as well as, suggesting a method for data augmentation that positively impacts BC prediction. Other work explores semi-supervised learning [3], offline RL [4], or temporal- and modality-attention modules [20]. Different from prior work, we are proposing a novel paradigm for learning social behaviors that does not aim to replicate behaviors from human-human interactions but allows the robot to stay socially appropriate during exploration for RL.

RL in HRI. There is significant research on RL applied to HRI. For example, prior work explored how humans can teach efficiently [21], [22], [23], or better align to baseline tasks [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], [29] to comfort, provide entertainment, and improve learning outcomes [30].

We focus on learning highly reactive robot behaviors. In this sense, our work relates to that of Qureshi et al. [5], who used online RL to learn a policy for a humanoid robot to interact with bypassing strangers. Different from our work, they use an internal mechanism to detect inappropriate behaviors by predicting reactions to the robot’s behavior. In

addition, inappropriate behaviors may not be as detrimental since new strangers pass by regularly.

Other related work focusing on offline RL for HRI explored learning non-verbal behaviors that aim to increase engagement in HRI [4], [31]. Closest to our work is recent work by Gillet et al. [6] which explores the use of offline RL to learn gaze behaviors that could balance participation in interaction between one robot and two human group members. Different from our work, they deploy the behaviors on test data to choose models that act socially appropriately.

Shielding in RL. Various frameworks incorporate the notion of safety in RL by emphasizing human intervention as a key component, though they diverge in their perspectives on the nature and scope of such intervention. In [32], the authors advocate for the active involvement of human experts at different stages of the training process to avert catastrophic actions. The concept of shielding was introduced by [2], employing linear temporal logic to establish high-level constraints that translate into sets of safe actions. In HRI, the application of shields has extended to discrete state and action spaces, like in cooking tasks for action modification [33], and to continuous state and action spaces in scenarios such as social navigation [34]. This work, to the best of our knowledge, is the first that designs continuous shields for safe backchanneling behavior.

III. PROBLEM DESCRIPTION

To formalize the use of shields for appropriate listening behaviors, we consider reinforcement learning setups with continuous state spaces $\mathcal{S} \subseteq \mathbb{R}^n, n \in \mathbb{N}^+$, and discrete action spaces $\mathcal{A} \subset \mathbb{N}^m, m \in \mathbb{N}^+$. To restrict the robot’s actions to those that can be safely executed at timestep t , we employ pre-shielding [2], which restricts the robot’s actions before they are applied. We model shields as functions $\mathcal{S} \rightarrow \text{pow}(\mathcal{A}_{\text{safe}})$ (pow denotes the power set), mapping the robot’s state to a set of safe actions. The set of safe actions, $\mathcal{A}_{\text{safe}}(s) := \Psi(s) \subseteq \mathcal{A}$, is obtained. More concretely, in our backchanneling application, shields Ψ restrict the robot’s action to a subset, $\mathcal{A} = \{\text{BCUt}, \text{Nod}, \text{Nod+BCUt}, \text{Do nothing}\}$ where BCUt are backchannel utterances, e.g., *mmmhh*.

IV. CREATING A SHIELD FOR SOCIALLY APPROPRIATE LISTENING BEHAVIOR

The shield Ψ for the reinforcement learning robot is built through a regression model bc_{model} and a corresponding threshold. This means that the regression model is trained to output a continuous value $o_t \in [0, 1]$ indicating how appropriate it is to backchannel at a given timestamp t . The threshold θ then shields actions as follows:

$$\Psi = \begin{cases} a_t = \text{Do nothing}, & \text{if } o_t \leq \theta \\ a_t \in \mathcal{A}, & \text{otherwise} \end{cases} \quad (1)$$

with \mathcal{A} being the entire set of actions after shielding.

A. Data preparation

We chose Cardiff’s Conversation Database (CCDb)[35] as our dataset to train the regression model. The CCDb provides unscripted, non-topic-bounded interactions that demonstrate a variety of listening behaviors and is publicly available.

The dataset consists of 30 dyadic conversations, each lasting around five minutes. The 30 conversations are between 16 different speakers (12 M, 4 F), with ages ranging from 25-56 years old. We extracted data from the perspective of each participant, totaling 115 minutes of conversational data.

We collected individual feature data for each participant separately. Since we are extracting data for listening behavior, we use the audio stream of the speaker as the input features and the annotated backchannels performed by the listener as the ground truth for the shield.

Shield input: As input features for our model, we extracted speech features from the speaker, including 13-dimensional mel-frequency cepstrum coefficients (MFCC) and 4-dimensional prosody features as used in prior works [19], [9], [3]. The MFCC features are computed every 30 ms with a sliding hamming window of 400 ms. Prosody features include pitch (fundamental frequency) and yin-energy, as well as the first derivative of these variables. The final 36-item feature vector is composed of the mean and standard deviation of each of these features (34 items) as well as one item for indicating active speech by the speaker and one item representing the robot state, i.e., if the robot backchanneled in the previous time step. We normalize the prosody data (34 items) for each speaker based on the first 30 seconds of data. Afterward, we normalize the entire dataset for training. In this work, we sample the features with 10Hz.

Regression model output: We focus on the timing of a backchannel to indicate if a backchannel is socially appropriate or not. To train the regression model $b_{C_{model}}$, we use the occurrence of a backchannel as an example for the maximum value of 1 and all other moments as the minimum value of 0. We further use the ground truth 700 ms earlier than the backchannel was annotated dataset, creating a gap between the prediction of backchannel and the need for execution. This gap to the annotation is based on the pause between signal and execution considered by [15] and allows us to accommodate a command-to-execution delay of the robot when producing backchannels.

B. Training a regression model as a shield for socially-appropriate listening behavior

To train sufficiently good regression models, we used six-fold cross validation to explore different model architectures with techniques such as data augmentation [9]. We upsampled the minority class to compensate for the imbalance of classes in the dataset. Below, we explain the models tested, evaluation criteria, and the selection process. We report model performance in Section V-A.2.

Hyperparameters: We explored two different architectures that have shown promise for backchannel generation in previous work [18]: Long Short-Term Memory (LSTM) and

Gated Recurrent Unit (GRU). During the training process, we explored a lookback, i.e. length of the time series fed into the model, of 2s, 4s and 8s, different optimization techniques (SGD, Adam), activation functions (sigmoid, relu), loss functions (mean squared error, mean absolute error, huber loss, and log cosh), batch sizes (8, 16, 32), dropout percentages (0, 0.2, 0.4) and different numbers of hidden units (8, 16, 32) of the recurrent networks in a grid search.

Data Augmentation: Murray et al. [9] suggest a method to improve the robustness of robot listening behaviors by using audio data augmentation. The authors show that the model trained on this data outperforms the models trained without data augmentation which was validated by [18], [36]. We emulated the proposed method by making use of *masking* techniques in the time and frequency domains. In our work, training instances (audio features) were partially masked in one or both domains, chosen at random. Both the original and the augmented instances were used for training. While upsampling, we augmented the samples independently of the original sample so that the same sample could be included in the dataset with different augmentations.

Threshold computation: The threshold θ is a key hyperparameter of the $\Psi(s_t) \rightarrow \{a_t^1, \dots, a_t^k\} \in \mathcal{A}_{safe}$ as it transforms the output o_T of the $b_{C_{model}}$ to the set of allowed actions (see Equation 1). We compute the threshold on the validation set. First, we observe how many backchannels appear in the validation set. The goal is to give the robot the opportunity to choose to backchannel 3.5 more times than are actually present in the validation set; allow a backchannel in every BOP. Therefore, we choose the lowest value as a threshold that leads to the shield allowing the full set of actions $a \in \mathcal{A}$ 3.5 times more than backchannels in the dataset.

Evaluation criteria: We use the recall of the backchannel class, i.e., how often the shield allows the total range of actions \mathcal{A} at time steps when the ground truth generated a backchannel. We first transform the output o_t of the regression model by replacing values with 1 if $o_t > \theta$, else 0. The threshold computation (see above) results that the output on the validation set contains 3.5 times the amount of values larger than the threshold than in the dataset. At the same time, the goal is that the shield allows backchannels in moments when the dataset had backchannels. For the measurement of recall, we allow a margin of 250ms for calculating the recall as suggested by [18]. This means we check a range of timesteps corresponding to 250ms before and after the backchannels to detect the correctness of the output. However, models with multiple repeated outputs above the threshold would largely benefit from this allowed margin. Therefore, we remove repeated BOPs in the output to ensure that models that generate distinct BOPs receive higher recall values than models that generate clusters of high values. If two models have the same recall value for the backchannel class (transformed value of 1), we choose the model with the higher macro recall value.

V. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of our shielding approach, we conduct a user study evaluating the appropriateness of a robot using a shielded random exploration approach.

A. User study - Evaluating the exploring RL robot

The user study aims to evaluate a robot using a shielded exploration strategy (SH) in comparison to two variants of unshielded exploration - one statistically guided unshielded approach (SG) and one completely randomly exploring approach (RA). We evaluate these three conditions in an online video study as visualized in Figure 1. We kept the input audio from the speaker the same in all three conditions but recorded different robot behaviors.

1) *Hypotheses*: For the experimental evaluation in the user study, we formulate three hypotheses covering the perception of the robot’s listening behavior, the appropriateness of backchannels, and the robot’s perceived intelligence.

As the shield was trained on a human-human dataset and aims to only allow backchannels in BOPs, we hypothesize:

- H1* The randomly exploring robot is perceived as a better listener with shielding than without shielding, i.e., when it is using a statistically guided (SG) exploration (*H1a*), or when it is using a random exploration method (*H1b*).
- H2* The robot is perceived as backchanneling more appropriately with shielding than without shielding, i.e., when it is using a statistically guided exploration (*H2a*), or when it is using a random exploration method (*H2b*).

As a result of the effects predicted in the first two hypotheses, we further hypothesize that backchanneling more appropriately and being a better listener might affect aspects of the robot’s perceived social intelligence. We hypothesize:

- H3* The robot is perceived as more socially competent and trustworthy, friendlier, and less rude while using the shield than without the shield, i.e., while it is using a statistically guided (SG) exploration (*H3a*), or when it is using a random exploration method (*H3b*).

2) *Conditions*: The user study used three conditions to evaluate the effectiveness of shielding for the exploring robot. Note that the input audio is exactly the same for all three conditions and was extracted from the *Talking with Hands* dataset [37]. Since we focused on socially appropriate timing, the robot randomly¹ decided which action of the full action set \mathcal{A} to use, i.e., if it used a backchannel utterance only (BCUtt), a nod only (Nod), noded and uttered a backchannel (BCUtt+Nod), or did nothing (Do nothing) whenever the shield allowed \mathcal{A} .

SH The shielded condition used a shield trained, evaluated, and selected as described in Section IV. The performance of the top three models is reported in Table I top and the final model’s performance on the top three folds in Table I bottom for each LSTMs (a) and GRUs (b) separately. The individual best-performing model was m_602, which was chosen for this condition. The

input audio stream was processed the same way as described in Section IV-A- *Regression model input* and then passed to the regression model $b_{C_{model}}$. With the help of the threshold θ , the shield decided which set of actions the robot may use in each time step according to equation 1. As the shield was allowing 3.5 times the amount of backchannels capturing the BOP, action Do nothing was chosen with a 3.5 higher likelihood than options BCUtt, Nod and BCUtt+Nod combined.

SG The statistically guided condition aimed to replicate approximately the same number of backchannels as the **SH** condition. The difference to **SH** was that the time steps at which backchannels were allowed were chosen randomly. Since we focused on replicating the number of time steps of executed notable backchannels from condition **SH**, action Do nothing was not used.

RA In this condition, the robot could use the full set of actions at every time step. To avoid continuous backchannel, action Do nothing was chosen with probability 0.5.

3) *Measures*: To evaluate our hypotheses, we conducted a video-based user study and used questionnaire-based measures to assess the effectiveness of the shielding approach.

Perceived Social Intelligence: We selected the four factors from the Perceived Social Intelligence Scale [38] that fit our experiment. We measured the robot’s **Social Competence (SOC)**, and how **friendly**, **rude**, and **trustworthy** the robot was perceived on a seven-level Likert item.

Appropriateness of backchannels: To specifically ask about backchanneling timing, we added three questions inspired by [9] asking whether the timing was appropriate, whether the timing was inappropriate, and if opportunities for backchannels were missed. Answers were given on a 5-level answer scale ranging from ‘Never’ to ‘Always’.

Listening quality: We adopted the scale proposed by Murray et al. [9] to measure the **perceived listening skill** and **feeling of comfort and closeness** under the premise of listening behavior measured on a seven-level Likert item. Like the original work, we found good reliability of the two factors measured through Cronbach’s alpha (perceived listening skill: $\alpha = 0.824$, feeling of comfort and closeness: $\alpha = 0.782$).

4) *Procedure*: After giving informed consent, participants answered demographic questions about their age, gender identity as well as participation in previous social robot user studies. Afterward, they were shown a video² of an interaction between the robot Furhat and a person. The person was shown from the back. Figure 1 shows a frame from this video. We used two audio snippets from session 32, take 18 (Minute 1:04-1:50, 5:31-5:54) from the *Talking with Hands* dataset [37]³ as the audio stimulus. We computed the voice scaler from session 32, 30 seconds of take 5. We

²Condition **SH**: <https://youtu.be/PemdDOE0xVc>, Condition **SG**: https://youtu.be/qsHJWxAU_Xw, Condition **RA**: <https://youtu.be/0IsNiGushF4>

³<https://github.com/facebookresearch/TalkingWithHands32M>

¹We set the random seed to 42.

TABLE I: Evaluation of different models by architecture allowing for a margin of 250ms. The top three rows for each architecture summarize evaluations over all folds and the lower three rows describe the best-performing individual shields.

(a) Models trained with LSTMs.

Identifier	Lookb.	Optim.	Act.	Loss function	batch s.	drop.	RNN u.	% full \mathcal{A}	Recall v.	Recall macro	θ
m_1464-m_1469	20	SGD	relu	huber loss	8	0	16	0.032	0.94	0.959	0.38
m_5784-m_5789	80	Adam	relu	mean squared error	32	0	16	0.035	0.936	0.956	0.476
m_2442-m_2447	20	Adam	relu	log cosh	8	0	32	0.036	0.936	0.956	0.209
m_1464	20	SGD	relu	huber loss	8	0	16	0.033	0.96	0.969	0.333
m_1466	20	SGD	relu	huber loss	8	0	16	0.027	0.955	0.967	0.496
m_1469	20	SGD	relu	huber loss	8	0	16	0.032	0.947	0.963	0.364

(b) Models trained with GRUs.

Identifier	Lookb.	Optim.	Act.	Loss function	batch s.	drop.	RNN u.	% full \mathcal{A}	Recall v.	Recall macro	θ
m_600-m_605	20	Adam	relu	mean squ. error	32	0	16	0.035	0.939	0.958	0.532
m_5784-m_5789	80	Adam	relu	mean squ. error	32	0	16	0.032	0.935	0.956	0.527
m_2280-m_2285	20	Adam	sigmoid	log cosh	8	0	32	0.032	0.932	0.954	0.498
m_602	20	Adam	relu	mean squ. error	32	0	16	0.033	0.96	0.97	0.604
m_603	20	Adam	relu	mean squ. error	32	0	16	0.041	0.952	0.961	0.436
m_600	20	Adam	relu	mean squ. error	32	0	16	0.03	0.944	0.96	0.672

mutated the listener’s voice to obtain the speaker’s audio only. To ensure an equal audio stimulus in all conditions, we used a boombox in front of the person to play back the audio. Unfortunately, the playback through the boombox led to reduced audio quality. Therefore, we showed subtitles to every participant. After watching the video, participants were asked to fill out attention check questions and the perceived social intelligence, listening behavior and appropriateness questions in the given order. Participants received 1.60€ compensation for participating in the study.

5) *System implementation*: The system ran with the help of the Robot Operating System ⁴ and the Furhat⁵ robot.

The system could run with any offline or online audio data. For the purpose of the study, we extracted the audio files from the *Talking with Hands* dataset [37]³ as described in Section V-A.4. From the audio files, we prerecorded rosbags with features sampled at 10Hz. For the final video recording, the same rosbag was used for generating the three different conditions. The decision-making was triggered based on the incoming audio features, which means that decisions were made at the same time stamps for all three conditions.

As there is no robot speech other than backchannels in the video stimuli, Furhat’s voice was set to the German voice ‘Andreas’ due to the sound of the backchannels. We randomly choose between ‘#MMM02#’ and ‘#MMM01#’ for backchannel utterances. The nod was implemented as a disposition of the neck randomly between 5 and 12 degrees. The face was set to ‘Titan’ as we wanted to keep the robotic appearance of the robot. We used the Furhat remote-api to control the robot. The shield’s main role was to ensure the appropriate timing of backchannels at 10Hz. However, when the robot queues commands, the delay between the command being sent and the robot executing is varying and sometimes it might take up to a few seconds. To avoid unpredictable delays in the system, we blocked sending commands to the

TABLE II: Age distribution of participants.

18-24	25-34	35-44	45-54	55-64	65+
7	34	22	15	8	6

robot for 800 ms after sending a command. In our system, we observed a delay from receiving the audio signal to hearable/visible command execution of ~ 200 ms which is within the 700 ms anticipated during data preparation. We delayed all commands by an additional 500 ms to match the training data. For all random choices, the seed was set to 42 at the beginning.

a) *Implementation of SG condition*: As probability-based decision-making can result in a randomly lower actual number of backchannels, we planned backchanneling on a fixed horizon, i.e. 30 seconds, ahead and distributed the backchannels randomly according to the given percentage of time steps that should have backchannels along the horizon. If only 20% of the planned time steps on the horizon were left, we planned the actions for the next horizon. After inspecting the recording, which contained the executed backchannels corresponding to the **SH** condition, we set the percentage to 4.4%, which was 26 backchannels in 590 time steps, i.e. 59 seconds.

6) *Participants*: We recruited in total 110 participants through Prolific⁶ of which 92 were used for the final analysis (46 female, 44 male, 1 non-binary, age distribution reported in Table II). 66 participants indicated having never participated in a social robotics study. 13 indicated they had participated in a social robotics study and 13 were unsure. We used Prolific’s recruitment procedure to only allow UK participants to avoid potential confounds of culture in our study. We ran an a priori power analysis in R⁷ for mixed Analysis of Variance (ANOVA) using an effect size of $f = 0.4$, an alpha of 0.05 with power set to 0.9. The

⁴<https://www.ros.org/>

⁵<https://furhatrobotics.com/>

⁶<https://www.prolific.com/>

⁷We used the pwr package in R 4.3.2

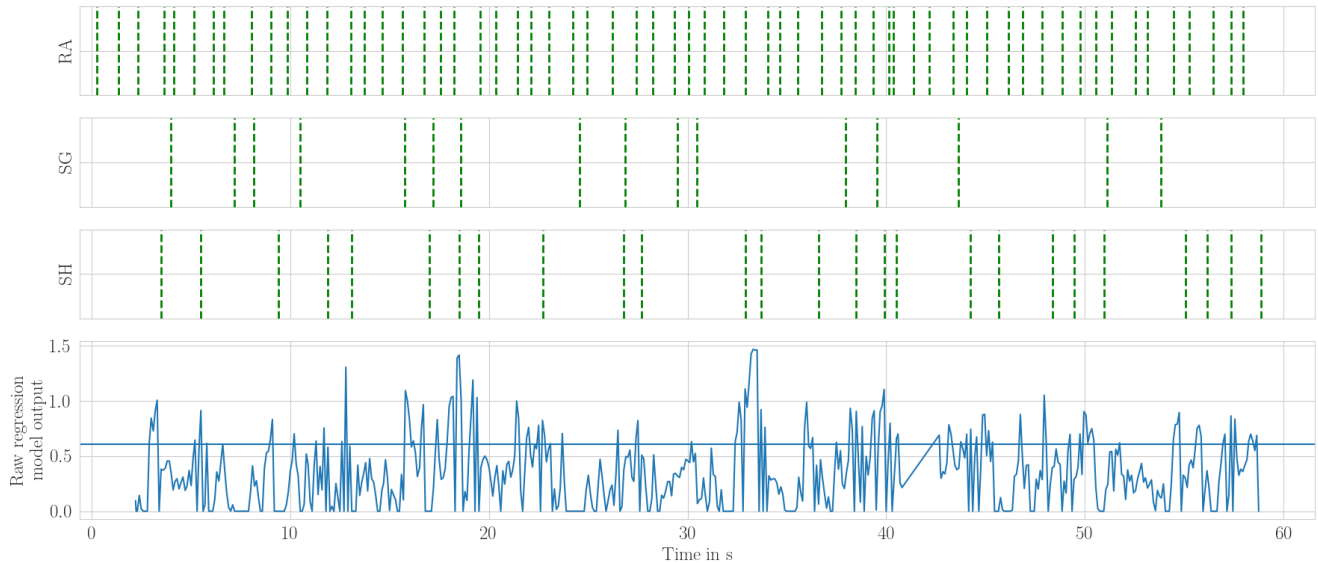


Fig. 2: Regression model outputs (blue solid) plotted with the respective threshold (horizontal line) (lowest plot). Above, green dashed lines indicate the moments in which the robot decided for an action other than `Do nothing` in the SH condition (lower middle), the SG condition (higher middle), and RA condition (top). The empty start is the time needed to collect the sequential data for the regression model, i.e., shield output.

power analysis suggested a total sample of 84 participants, meaning 28 participants for each of the three conditions. After the initial experiments, we ran subsequent experiments to compensate for excluded participants until one condition had at least 28 participants. 16 participants were excluded from the analysis as they failed the first (10 participants) or second (6 additional participants) attention check. Attention checks asked for video content and fixed ratings on one question.

B. Results from user study

Perceived Social Intelligence: After verifying normality and homogeneity of variance, we performed a one-way ANOVA on the friendly scale. We did not find a significant influence of condition on how friendly participants thought the robot was ($F(2) = 0.29, p = 0.744$). The remainder of the outcome variables for perceived intelligence were non-normal. A Kruskal-Wallis H test showed that there was a statistically significant difference in how rude the robot was perceived between the different conditions, $\chi^2(2) = 9.12, p = 0.01$, with a $M = 3.23, SD = 0.77$ for SH, $M = 2.71, SD = 0.98$ for SG and $M = 3.56, SD = 1.33$ for RA. A pairwise comparison with Bonferroni correction showed that the RA condition ($p = 0.025$) and the SH condition ($p = 0.046$) were perceived as significantly more rude than the SG condition. The other outcome variables measuring trustworthiness ($\chi^2(2) = 1.75, p = 0.415$) and social competence ($\chi^2(2) = 2.13, p = 0.345$) were not significantly different between conditions.

Appropriateness of backchannels: As the three additional questions about appropriateness did not show reliable overlap ($\alpha = 0.4$), we evaluated each question separately. As the data was non-normal, we used a Chi-squared test with condition

as the predictor. We found that condition significantly affected the amount of appropriate timing $\chi^2(8) = 30.334, p < 0.001$. A pairwise comparison with Bonferroni correction revealed that there was more appropriate backchanneling in the SG condition compared to the RA condition ($p < 0.001$) and SH condition compared to the RA condition ($p = 0.038$) with $M = 3.36, SD = 0.82$ for SG, $M = 3.39, SD = 0.73$ for SH, and $M = 2.94, SD = 1.12$ for RA. Similarly, the amount of inappropriate timing was affected by condition $\chi^2(8) = 26.556, p < 0.001$. A pairwise comparison with Bonferroni correction revealed a significant difference between the SG and RA condition ($p < 0.001$) with $M = 2.33, SD = 0.92$ for SG, $M = 2.89, SD = 1.06$ for SH, and $M = 3.74, SD = 1.09$ for RA.

Listening quality: Both factors measuring listening quality were not normally distributed. A Kruskal-Wallis H test showed that there was a statistically significant difference in how comforting and close the robot was perceived between the different conditions, $\chi^2(2) = 6.924, p = 0.031$, with $M = 3.68, SD = 1.08$ for SH, $M = 3.28, SD = 1.1$ for SG and $M = 3.00, SD = 1.12$ for RA. A pairwise comparison with Bonferroni correction revealed a significant difference between the RA and SH condition $p = 0.026$. The perceived listening skill was not significantly different between conditions ($\chi^2(2) = 5.44, p = 0.065$).

C. Model performance in scenario

The *Talking with Hands* dataset used to extract audio for the user study does not provide annotations. Thus, we cannot evaluate the objective quality of the backchannels outside of the user study. Figure 2 shows the output of the regression model, the threshold, and the respective random decision by

shield SH. For comparison, we further plotted the output of the SG and RA condition, which we discuss in Section VI.

VI. DISCUSSION & LIMITATIONS

In this work, we evaluated if a shielding approach would allow for socially appropriate behavior during the exploration phase in an RL setting. We compared shielded exploration with two other exploration methods: one that represents a feasible and straightforward approach to exploration (RA), and another exploration method that offers a more thoughtful comparison that still offers valid exploration (SG). Based on the results from the video study (refer to Sec. V-B), we did not find evidence in favor of the shielded condition (SH) when compared to the statistically guided exploration (SG). Therefore, we have to reject all hypotheses—H1a, H2a, H3a—concerning the comparison between these two conditions.

The chosen random seed for the experimental evaluation might have affected the results for the SG condition and thus the comparison between the SH and SG conditions. We set the seed prior to running any of the recordings for the user study. In Figure 2, we can observe that there were fewer backchannels present in the SG condition despite the careful implementation to ensure similar amounts of backchannels (see Section V-A.5). One reason for this smaller number of backchannels is that the randomly distributed backchannels appeared close to each other which led to them being filtered out by the execution system (see Section V-A.5). Further, a portion of the executed backchannels in the SG condition were by chance close to BOPs as identified by our shield (bottom plot in Figure 2). It is possible that another random seed would have resulted in different experimental outcomes. Future work should therefore consider evaluations on a larger range of stimuli. For the RA condition, we do not expect that the outcome would change with a different seed due to the almost continuous backchanneling.

Despite the continuous backchanneling in the RA condition, we have to reject H3 as we did not find significant differences for perceived intelligence in favor of the shielded exploration. Interestingly, the SG condition was perceived as significantly less rude than both the SH and RA condition. As discussed above, this result could be a result of by chance appropriate timing or due to the smaller number of backchannels present in the SG condition (see Figure 2).

Despite the lack of difference for perceived intelligence, we can partially accept H1b and H2b indicating that the robot in the SH condition was perceived as closer and more comforting (H1b) and as more appropriate (H2b) than the robot in the RA condition. Interestingly, the perceived listening skill was not affected.

Based on the results from our user study, we can conclude that the shielded exploration method (SH) or statistically guided exploration (SG) might provide sufficient interaction quality that would allow for reinforcement learning. However, we expect that the statistically guided exploration would take longer to explore the relevant state space due to the scarcity of backchannels and the lack of any additional guidance such as the shield.

One limitation of our work is the limited variability of the interaction context in the dataset and the dataset size. Future work should explore the effects of the amount and variability of data on the effectiveness of the shields and explore how large datasets for listening behaviors could be collected efficiently.

We note that our results might not be completely realistic as it is unlikely that a speaker would not react to extensive or falsely timed backchannels. To ensure that participants' perceptions in the user study were solely based on the alteration of the robot's behavior, we kept the audio stimulus constant. This approach allowed us to remove possible confounds based on the speaker's reactions to the robot's listening behavior. However, one potential reason for the lack of significant difference between conditions for some of the measures could be due to the lack of adaption of the human speaker to the potentially mistimed backchannels. In this study, we decided to focus on the robot's behavior and avoid confounds due to speaker reactions. Future work will need to further investigate the promise of shielding for socially appropriate listening behavior in interaction especially given the robot's behavior potentially altering the human's reaction in the exploration phase.

Backchannels and other communicative behaviors are idiosyncratic [13], i.e., they are peculiar to an individual. For robots to develop distinct communicative behaviors, techniques such as reinforcement learning might be more suitable than, imitation learning which would merely replicate human behavior. This work provides promising indications that shielding could be one approach to ensure appropriate backchanneling behavior during exploration. Future work will need to explore if statistically guided approaches can be leveraged or if shielding has additional benefits, e.g. needing fewer data points. In this case, the effort of training and choosing a data-driven shield might be favorable considering the potential benefits of fewer data points for online training with human participants interacting with the system.

VII. CONCLUSION

This work proposes shielding for RL as a novel paradigm to learn idiosyncratic social behaviors, e.g., listening behavior. We provide a full problem formulation and focus on studying the exploration phase of the RL problem. We compare the proposed shielding approach to two other random exploration methods - one statistically guided approach and one completely random approach. In a video-based user study, we show that the shielded and statistically guided exploration approaches are perceived as having higher listening quality, being more appropriate (shield) or being less rude and less inappropriate (statistically guided). Therefore, these two approaches are better suited for creating sensible interactions during the exploration phase than a fully random approach to exploration. Future work will need to answer if both approaches are similarly suitable for efficient learning of listening behavior in interaction with people.

REFERENCES

- [1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [2] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, “Safe reinforcement learning via shielding,” in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [3] 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.
- [4] N. Hussain, E. Erzin, T. Metin Sezgin, and Y. Yemez, “Batch recurrent q-learning for backchannel generation towards engaging agents,” *arXiv*, 2019.
- [5] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, “Intrinsically motivated reinforcement learning for human–robot interaction in the real-world,” *Neural Networks*, vol. 107, 2018.
- [6] S. Gillet, M. T. Parreira, M. Vázquez, and I. Leite, “Learning gaze behaviors for balancing participation in group human-robot interactions,” in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '22. IEEE Press, 2022.
- [7] V. H. Yngve, “On getting a word in edgewise,” in *Chicago Linguistic Society, Chicago*, 1970.
- [8] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency, “Virtual rapport,” in *Intelligent Virtual Agents*, J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- [9] M. Murray, N. Walker, A. Nanavati, P. Alves-Oliveira, N. Filippov, A. Sauppe, B. Mutlu, and M. Cakmak, “Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations,” in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022.
- [10] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen, “Backchannel strategies for artificial listeners,” in *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, ser. IVA'10. Berlin, Heidelberg: Springer-Verlag, 2010.
- [11] P. Blomsmma, G. Skantze, and M. Swerts, “Backchannel behavior influences the perceived personality of human and artificial communication partners,” *Frontiers in Artificial Intelligence*, vol. 5, 2022.
- [12] M. Heldner, A. Hjalmarsson, and J. Edlund, “Backchannel relevance spaces,” in *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*. Peter Lang Publishing Group, 2013.
- [13] P. Blomsmma, J. Vaitonyté, G. Skantze, and M. Swerts, “Backchannel behavior is idiosyncratic,” *Language and Cognition*, 2024.
- [14] K. Truong, R. Poppe, and D. Heylen, “A rule-based backchannel prediction model using pitch and pause information,” in *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA), Sep. 2010, null ; Conference date: 26-09-2010 Through 30-09-2010.
- [15] N. Ward and W. Tsukahara, “Prosodic features which cue back-channel responses in english and japanese,” *Journal of Pragmatics*, vol. 32, no. 8, 2000.
- [16] Y. Okato, K. Kato, M. Kamamoto, and S. Itahashi, “Insertion of interjectory response based on prosodic information,” in *Proceedings of IVTTA '96. Workshop on Interactive Voice Technology for Telecommunications Applications*, 1996.
- [17] L.-P. Morency, I. Kok, and J. Gratch, “A probabilistic multimodal approach for predicting listener backchannels,” *Autonomous Agents and Multi-Agent Systems*, vol. 20, 01 2010.
- [18] M. T. Parreira, S. Gillet, and I. Leite, “Robot duck debugging: Can attentive listening improve problem solving?” in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, 9 2023.
- [19] R. Ruede, M. Müller, S. Stüker, and A. Waibel, *Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor: 8th International Workshop on Spoken Dialog Systems*, 01 2019.
- [20] K. Wang, M. M. Cheung, Y. Zhang, C. Yang, P. Q. Chen, E. Y. Fu, and G. Ngai, “Unveiling subtle cues: Backchannel detection using temporal multimodal attention networks,” in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023.
- [21] 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, 2006.
- [22] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, “Variquery: Vae segment-based active learning for query selection in preference-based reinforcement learning,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.
- [23] J. de Heuvel, F. Seiler, and M. Bennewitz, “Enquery: Ensemble policies for diverse query-generation in preference alignment of robot navigation,” *arXiv preprint arXiv:2404.04852*, 2024.
- [24] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, “Aligning human preferences with baseline objectives in reinforcement learning,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.
- [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, 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.
- [27] K. Weber, H. Ritschel, I. Aslan, F. Lingensfelder, 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*, 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, 2017.
- [29] S. Holk, D. Marta, and I. Leite, “Predilect: Preferences delineated with zero-shot language-based reasoning in reinforcement learning,” in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024.
- [30] 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, 2016.
- [31] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, “Training socially engaging robots: Modeling backchannel behaviors with batch reinforcement learning,” *IEEE Transactions on Affective Computing*, 2022.
- [32] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, “Trial without error: Towards safe reinforcement learning via human intervention,” in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018.
- [33] S. Van Waveren, C. Pek, J. Tumova, and I. Leite, “Correct me if i’m wrong: Using non-experts to repair reinforcement learning policies,” in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022.
- [34] D. Marta, C. Pek, G. I. Melsión, J. Tumova, and I. Leite, “Human-feedback shield synthesis for perceived safety in deep reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 7, no. 1, 2021.
- [35] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendevanter, D. W. Cunningham, and C. Wallraven, “Cardiff conversation database (ccdb): A database of natural dyadic conversations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.
- [36] M. T. Parreira, S. Gillet, K. Winkle, and I. Leite, “How Did We Miss This? A Case Study on Unintended Biases in Robot Social Behavior,” in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 3 2023.
- [37] G. Lee, Z. Deng, S. Ma, T. Shiratori, S. S. Srinivasa, and Y. Sheikh, “Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [38] K. A. Barchard, L. Lapping-Carr, R. S. Westfall, A. Fink-Armold, S. B. Banisettey, and D. Feil-Seifer, “Measuring the perceived social intelligence of robots,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 4, 2020.