Training Socially Engaging Robots: Modeling Backchannel Behaviors with Batch Reinforcement Learning

Nusrah Hussain, Engin Erzin, *Senior Member, IEEE*, T. Metin Sezgin, and Yücel Yemez

Abstract—A key aspect of social human-robot interaction is natural non-verbal communication. In this work, we train an agent with batch reinforcement learning to generate nods and smiles as backchannels in order to increase the naturalness of the interaction and to engage humans. We introduce the Sequential Random Deep Q-Network (SRDQN) method to learn a policy for backchannel generation, that explicitly maximizes user engagement. The proposed SRDQN method outperforms the existing vanilla Q-learning methods when evaluated using off-policy policy evaluation techniques. Furthermore, to verify the effectiveness of SRDQN, a human-robot experiment has been designed and conducted with an expressive 3d robot head. The experiment is based on a story-shaping game designed to create an interactive social activity with the robot. The engagement of the participants during the interaction is computed from user's social signals like backchannels, mutual gaze and adjacency pair. The subjective feedback from participants and the engagement values strongly indicate that our framework is a step forward towards the autonomous learning of a socially acceptable backchanneling behavior.

Index Terms—Human-robot interaction, user engagement, backchannels, batch reinforcement learning

1 Introduction

Social robots are becoming more integrated into our daily lives, with an increasing prevalence in human environments such as schools, museums and hospitals. Contrary to industrial or service robots, these robots work in close proximity to humans and are expected to display social behaviors that encourage their acceptance in the human company [1]. From a broad perspective, the design of a social robot aims a human-robot interaction (HRI) that is perceived similar to a human-human interaction. Several design characteristics contribute to achieving this objective such as advanced conversation skills, emotion recognition and display capabilities, user response based behavior adaptation, ability to develop a social relationship and the potential of displaying varying social roles [2].

An important characteristic desired from a social robot is its ability to sustain user engagement [3], [4]. Many have set engagement as a common goal of human-robot interaction and a metric to gauge its success [5], [6]. HRI with objectives such as teaching a new skill, guiding in a public place, and aiding in physical therapy, are some examples where engagement is a key design feature. An attentive and engaged user is more likely to benefit from the service compared to disengaged one. However, engagement is a complex concept, with numerous proposed definitions [7]. It is an interdisciplinary field between social sciences and robotics [8]. While social scientists try to understand behaviors among humans that enhance engagement, robotic engineers aim to replicate these behaviors in a robot.

A key strategy that humans use to engage their interlocutors during interaction is through verbal and non-verbal cues. While the generation of such cues by the speaker is of great importance [9], [10], [11], the generation of similar feedback signals known as backchannels by the listener is equally important [12], [13], [14], [15], [16]. Backchannels are non-intrusive social signals generated during the listening turn [17]. They include facial expressions, body gestures, display of emotions, and verbal expressions such as 'yeah' and 'hmms'. Several empirical works, in a human-robot interaction setup, have found that users are more engaged when interacting with a backchanneling robot [18], [19]. The common approach for automation of backchannel behavior in robots is using rule-based methods. Supervised methods can also be applied if a dataset is created with optimum behavior demonstrations. However, these approaches do not bring a sense of purpose explicitly into the robot (i.e., to engage user). Moreover, the behavior learned from supervised learning cannot perform better than the reference behavior of the dataset. Recent works have demonstrated the use of online reinforcement learning (RL) for engagement maximization. These works incorporate user's social signals to measure engagement and exploit it as the reward of the RL algorithm [20], [21]. A major issue faced by the online methods is the necessity of direct interaction

- *The authors are with the KUIS AI Lab., College of Engineering, Koç University, 34450 Istanbul, Turkey. E-mail: {nhussain15, eerzin, mtsezgin, yyemez}@ku.edu.tr.*

Manuscript received 16 November 2021; revised 24 May 2022; accepted 29 June 2022. Date of publication 13 July 2022; date of current version 15 November 2022.
This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 217E040. The work of Nusrah Hussain was supported by Higher Education Commission (HEC) Pakistan.
(Corresponding author: Nusrah Hussain.)
Recommended for acceptance by J. Broekens.
Digital Object Identifier no. 10.1109/TAFFC.2022.3190233

with users during training. The training may be highly time-consuming and early interactions with an untrained robot can be frustrating.

We suggest the use of batch reinforcement learning (batch-RL) to train a robot for engaging behaviors in an offline manner. Batch-RL determines optimum solution for sequential decision-making problems using a batch of trajectories collected by other behaviors, such as an expert human [22], [23]. The available human-human interaction datasets represent behavior trajectories when a human interacts with another. With batch-RL techniques, these datasets may be exploited to create decision-making engines. In this work, we aim to train an agent for non-verbal behaviors (smiles and nods) as backchannels which maximizes user engagement. The human-human interaction dataset is processed for batch-RL, where the rewards come from user engagement. The nods and smiles (including visual and audible laughter in the dataset) are annotated as actions. The preliminary version of this approach was presented in our earlier works [24], [25]. Although our previous work on offline RL has shown promising results in terms of learning an engaging backchannel policy, off-policy methods used for offline RL are known to suffer from distributional shift [72]. While the function approximator is trained under one distribution, it might face a different distribution when interacting with the environment. This problem is particularly apparent in the case of backchannel learning since backchannel events are usually very sparse in human-human communication and datasets available are rather biased towards rewarded laughs with almost no demonstration of negative laughs and of what would happen for example when backchannels were used in excess. Hence off-policy RL methods are prone to learning policies that generate backchannels more frequently than a human would do.

This paper is an extension of our preliminary work with the following contributions.

- We propose the *Sequential Random Deep Q-Network (SRDQN)* as a batch-RL algorithm to train an agent for non-verbal gestures. The performance of the proposed method is evaluated with offline evaluation methods and compared to the existing batch-RL methods such as neural fitted Q-iterations (NFQ) [26] and deep Q-network (DQN) [27].
- We address the distributional shift problem by constraining the frequency of backchannels generated by the RL policy to remain close to the frequency demonstrated in the dataset. We achieve this by introducing a reward factor which penalizes the rewards that result in excessive number of smiles/nods.
- We train our RL agent for two types of backchannels: nods and smiles. Both backchannels are trained independently with the SRDQN algorithm and their performances are compared. Note that, our previous work considered only smile generation.
- We conduct human-robot interaction experiments with the trained RL-agent. The social robot Furhat is used to interact with human subjects in a story-shaping game. During the interaction, the robot generates nods and smiles as backchannels following its RL-policy. The results of the user engagement analyses and the feedback questionnaire favour the RL trained system over a baseline rule-based policy.
- An important inference from our user study is the lower acceptability by the users of untimely smiles compared to nods. This indicates that while nod policy can be more flexible, the smile policy needs to be learned closer to the optimal policy.

2 Related Work

Social robots can already be seen as tutors [28], guides in public places like airports and museums [29], assistants in work environments [30], healthcare robots for the elderly [31], [32], and facilitators for children with autism spectrum disorders [33]. For the central goal of user engagement, several approaches have been used in the design of backchannel behavior. The simplest one defines rule-based behaviors, where the decision to trigger a backchannel is conditioned on the user's reply. In a study with educational robots [34], a language learning experiment is conducted. The authors show that supportive expressions such as "don't worry" and encouraging behaviors like nod/smile at correct answers improve the learning performance of the students. Similarly, in another study with a robot, various backchannel strategies are used during a mathematics test conducted on a tablet device [35]. The authors show the effectiveness of non-verbal feedback (nod/shake) after each answer, gaze shift from student to tablet and the use of supportive phrases in improving engagement and test outcomes of the students. Besides tutoring, healthcare social robots for assistance and companionship of the elderly, are becoming increasingly popular [31]. One of the challenges with implementing and designing healthcare robot is its acceptance by the elderly. In a study with an assistive activity, the interaction experience of cognitively impaired seniors was investigated with a robot capable of dynamic facial expressions and gestures [19]. The results showed that a human-like robot having an expressive face and arm gestures significantly increases levels of engagement, positive affect, and perceived social intelligence during the interaction.

However, only certain behaviors can be generalized with a simple rule-based policy; the ones that are observed as typical among humans. Some behaviors differ significantly from person to person, and hence require more intelligent behavioral strategies [36]. The authors in [37] use sequential probabilistic models (e.g., Hidden Markov Models or Conditional Random Fields) with supervised learning to predict listener backchannels using the speaker multimodal output feature. In [38], a backchannel is triggered with some probability either when the user nods, or when a variation is observed in the pitch of the user's voice. The authors show that robots that generate multi-modal backchannels under a certain probabilistic rule, are perceived as sensitive listeners and are more competent in sustaining engagement and the conversational dialog. In another work, the prosody and pause behaviors of the user are used to construct a rule-based backchannel policy [39]. Engagement has also been shown to improve when the robot mirrors the laughs of user [40]. In [41], listener's head nods are generated based on a speaker-adaptive prediction model using a corpus of dyadic interactions, while the work in [42] investigates how gaze, in addition to prosody, can cue backchannels. There are also other works in the literature, which aim to learn backchannel behaviour from recorded human-human interaction datasets via supervised learning [43], [44], [45].

Recent works have been exploring reinforcement learning for the design of more intelligent behavior strategies. The strength of RL framework lies in the concept of rewards, which allows the goal of the interaction to be integrated into the formulation. In HRI, the robot behavior has been optimized for a wide range of objectives. Existing RL-based works focus on learning empathetic supportive strategies [46], affective behavior [21], human-like greeting approach [47] and natural interaction distance and gaze control [48], [49]. While RL is a popular learning framework in HRI, research works that incorporate engagement and related social signals as rewards are still scarce. In [20], online reinforcement learning is used to adapt the personality of a robot by varying extroversion through linguistic styles, with the goal of keeping the user engaged in a storytelling scenario . In this study, engagement is estimated using head tilt and openness of the body. In the work by Weber *et al.* [50], the robot's sense of humor is adapted to the user's preference as an approach to engage and bond with the user, where the reward signal comes from user's smiles and laughs as indicators of engagement. In a language tutoring setup [21], facial expressions is exploited to measure a child's engagement and use it to adapt robot's behaviors.

The success of online reinforcement learning is however limited in HRI due to the tediousness of interaction with human. Researchers are now looking for solutions with batch reinforcement (or offline reinforcement) learning for social robots. The potential of batch-RL in learning language and dialog has been demonstrated on corpora of agent-customer transcripts [51] and on collection of human-bot interaction data [52]. The popularity of batch-RL is rising in not only in HRI but also in numerous real-world applications such as safety-critical healthcare treatment [53], [54], risk-prone self-driving cars [55], [56], large-scale learning for recommender systems [57], [58], robot navigation [59], and grasping tasks [60]. QT-Opt [61] is described as a Q-learning algorithm that can learn effective vision-based robotic grasping strategies from hundreds of thousands grasping trials. Fitted Q-iteration is used in [62] on data collected from clinical trial involving 1460 patients to learn the optimum treatment options for schizophrenia. However, the use of batch-RL for engagement is yet to be explored. Batch-RL faces the key challenge of distributional shift when evaluated on unseen data outside of training corpus due to both the change in visited states for the learned policy and the act of maximizing the expected return [63]. A common way to address this problem is to impose constraints on the learning process. In [64], batch-constrained reinforcement learning is introduced, which restricts the action space so as to force the agent towards behaving close to on-policy with respect to a subset of the given data. In [65], conservative Q-learning (CQL) is proposed, which aims to learn a conservative Q-function such that the expected value of a policy under this Q-function lower-bounds its true value. Another method used is to add a penalty term to the reward function to avoid action decisions that would result in large deviations from the behavior policy of the dataset [66], [67].

To the best of our knowledge, this work (including our preliminary papers [24], [25]) is the first work on batch reinforcement learning for engaging backchannel behavior of a robot. We also note that a very recent work [68] uses conservative Q-learning as a batch-RL algorithm to learn a backchannel policy that enhances engagement while statistically matching the human laughter generation in dyadic conversations. It, however, only trains for laugh events and does not include any user study that validates the proposed method on a real human-robot interaction setting.

3 Methodology

We formulate the problem of backchannel generation in HRI as a Markov decision process (MDP), which we solve with offline batch-RL using a human-human interaction dataset as the batch of samples. In this section we describe our MDP formulation and the Sequential Random Deep Q-Network (SRDQN) algorithm proposed as a batch-RL method to address this problem.

3.1 MDP Formulation

Markov decision process is commonly used for stochastic control problems, and is given by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is action space, $\mathcal{P}$ is the state transition probability function, and $\mathcal{R}$ is the reward function. In our formulation, at time step t, the user (i.e., the environment) is in a state $s_t \in \mathcal{S}$, the robot takes a backchanneling action $a_t \in \mathcal{A}$ and a scalar reward (i.e., engagement) $r_t \sim \mathcal{R}(s_t, a_t)$ is generated by the user. At the next time step, the user transitions to a new state $s_{t+1} \sim \mathcal{P}(s_t, a_t)$. The discount factor $\gamma \in [0, 1)$ weighs the future rewards, determining the temporal extent of the action's effects. The optimal solution to the MDP is a backchannel policy $\pi(a|s)$ which maximizes the expectation of the sum of discounted rewards, i.e., the overall engagement in our case. The policy is defined only for listening turns of the robot when it provides feedback to the speaker (user) with backchannels. When learning the nod behavior, two actions are defined: 'do nothing' or 'nod'. Similarly, the actions 'do nothing' or 'smile' are described when training for smile behavior.

We represent the user state with a statistical summarization of Mel-Frequency Cepstrum Coefficients (MFCCs) and prosody features from the user's speech. The speech features are determined at a rate of 40 Hz. MFCCs are computed as a 13-dimensional vector using a 40ms Hamming window. Prosody features are defined as a 6-dimensional vector comprising of speech intensity, pitch and confidence-to-pitch, along with their first derivates. Both features are concatenated to create a 19-dimensional feature vector at every 25ms time step. The state of the process is defined from a history of speech features. Feature summarization is performed over one second of speech features (40 samples) by calculating 11 statistical measures over each dimension. These measures are mean, standard deviation, skewness, kurtosis, range, minimum, maximum, first quartile, third quartile, median and inter-quartile range [69]. The resultant 209-dimensional state feature describes the short-term distribution of the speech features. The optimum policy of the robot will define a nod (or smile) behavior as backchannel, conditioned on the user speech features, that maximizes the reward, i.e., user engagement.

3.2 Engagement Measurement

Engagement can be inferred objectively from various observable features and behavior cues [7]. The simplest methods adopt single variables as proxies of engagement such as smiles and laughs [50], or disengagement such as face location [70] and negative emotions [71]. The more common approach, however, is to use several verbal and non-verbal social signals that arise over a certain time window [72]. Posture, head tilt and orientation, backchannels and gaze are some of the cues used to measure engagement [20], [73]. In this work, we follow the method proposed by Rich *et al.* for measuring engagement in a dyadic face-to-face interaction [74]. The authors define four connection events (CE) to measure engagement: backchannels, adjacency pair, mutual facial gaze, and directed gaze. Backchannels are the feedback gesture events generated from user during listening turn. Adjacency pair event is triggered when speech turn-taking occurs with some minimal time gap. The mutual facial gaze occurs when there is face-to-face eye contact. Lastly, directed gaze event is defined when both participants look at a nearby object related to the interaction at the same time. Engagement $\mathcal{E}(t)$ is then defined as the average number of CEs over a time window as

$$\mathcal{E}(t) = \frac{\eta_t - \eta_{(t-\Delta)}}{\Delta}, \tag{1}$$

where η_t is the number of connection events occurred until time t and Δ is the time window size.

Similarly, we extract the connection events from our dyadic human-human interaction dataset. To describe backchannel events, we use smiles, laughs, nods and head shakes. Adjacency pair and mutual gaze events are also annotated over the dataset. Since in our dataset, we do not have objects of interest at which both parties look at, we exclude directed gaze event from our definition. Note that the reward r_t in our MDP formulation is given by $\mathcal{E}(t)$ which is calculated over a 15 seconds window size in our experiments to avoid abrupt changes in reward function since human engagement varies slowly over time.

3.3 Off-Policy Batch Data

Batch reinforcement learning algorithms utilize only previously collected data logged in the form of tuples (s_t, a_t, r_t, s_{t+1}). Such data-driven offline methods allow the manipulation of large datasets for sequential decision making problems. There exists no human-human interaction dataset that specifically aims to increase engagement by generating backchannels. Any sufficiently large dataset which contains a variability in engagement values with a range of annotations for determining the states, actions and rewards may be used as an offline batch of transitions. We work with the IEMOCAP dataset [75], which is designed to analyze expressive human interactions. It consists of five sessions acted by ten professional actors performing dyadic human-to-human conversations. In total, there are 151 dialogues performed in pairs of the opposite gender, on 8 hypothetical scenes and 3 scripted plays. This dataset represents human behavior in a variety of situations where the policy behind each dialog may vary. The nod and smile gestures (laughs are included in smiles), are annotated in the dataset[1]. The (s_t, a_t, r_t, s_{t+1}) tuples are prepared from the dataset in order to create an off-policy batch of trajectories. In each dialog between two actors, the backchannel actions (nods and smiles) are labeled from one actor and state (speech features) and rewards (engagement) are determined from the other. Thus, in terms of the RL scheme the first actor executes actions while second behaves like the interaction environment. Fig. 1 shows the relative time windows for the extraction of states, actions, and rewards. The dialogues are processed at 40 Hz and the consecutive 25ms time windows are labelled with 1 or 0 to indicate the presence or absence of the backchannel event a_t. The state of the user is described by a history of speech features from one second window preceding the action. The consequence of the action a_t in state s_t is followed with a reward r_t, determined over a 15 second time window.

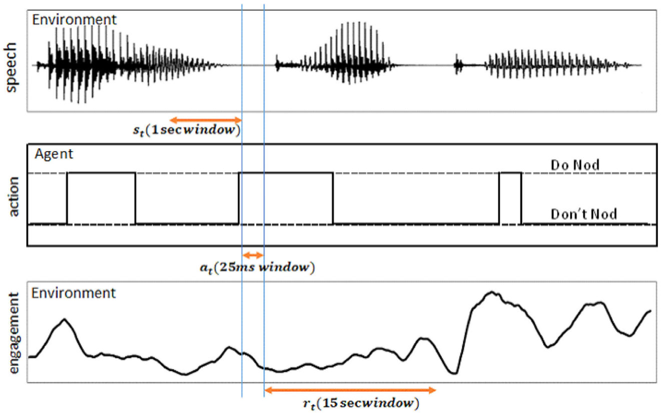


Fig. 1. Reinforcement learning formulation of speech driven backchannel generation (not drawn to scale).

3.4 Algorithm: Sequential Random Deep Q-Network (SRDQN)

We introduce the Sequential Random Deep Q-Network (SRDQN) as an off-policy batch-RL algorithm for learning a backchannel policy for robot during human-robot interaction. It belongs to the class of Q-learning algorithms where the distributions for rewards and transition dynamics are not explicitly learned. The quantity of interest is state-action value function, denoted $Q^{\pi}(s, a)$, which gives the expected future return starting from a particular state-action tuple. The goal of SRDQN is to learn the optimum Q-value function, $Q^{*}(s, a)$, using only previously collected transition samples (s_t, a_t, r_t, s_{t+1}) stored in an experience replay buffer $\mathcal{D}$.

SRDQN estimates the Q-values with a multi-layer perceptron function approximator, as proposed earlier by Riedmiller in the neural fitted Q-iteration (NFQ) algorithm [26]. SRDQN follows the iterative process of fitting the Bellman control equation to the samples in $\mathcal{D}$. The Q-values are

[1] Head nod and smile gesture annotations on the IEMOCAP dataset are available on https://mvgl.ku.edu.tr/databases/

TABLE 1
List of Hyperparamters and Their Values

Hyperparameter	Symbol	Value	Description
minibatch size	m	256	Number of training examples over which gradient descent is performed
replay memory size	C	4800	Capacity of $\mathcal{D}$ from which a minibatch is sampled
history length	H	1 sec	Duration of audio data used to define the states.
target network update frequency	N	100	Frequency with which target network is updated (measured in number of parameter updates)
discount factor	γ	0.99	Discount factor used in Q-learning update
learning rate	l_r	0.00005	Initial learning rate used by the optimizer
reward factor	κ	0.9	Factor for scaling rewards that are generated from backchannel actions
optimizer	-	Adam	Optimization algorithm
exponential LR scheduler	-	0.99	Exponential factor used to decay the learning rate per epoch
activation function	-	ReLU	Non-linearity between each linear layer
weight initialization	-	normal random	Initialization method of neural network parameters

estimated with neural network Q_θ, parameterized with weights θ. A separate network $\hat{Q}_{\theta'}$ determines the target value, parameterized by θ', that is periodically updated after every N gradient updates. The Bellman control equation then takes a form similar to that proposed by deep Q-network (DQN) algorithm [27], and is given by

$$Q(s_t, a_t; \theta) = r_t + \gamma \max_a \hat{Q}(s_{t+1}, a; \theta'). \tag{2}$$

We make the standard assumption that the future rewards are discounted by a factor of $\gamma = 0.99$ per time-step. Given the target value $y_t = r_t + \gamma \max_a \hat{Q}(s_{t+1}, a; \theta')$ and the error $e = Q(s_t, a_t; \theta) - y_t$, the loss function $\mathcal{L}$ is defined by the smooth L1 loss function (i.e., Huber loss) as

$$\mathcal{L} = \begin{cases} 0.5e^2, & \text{if } |e| < 1 \\ |e| - 0.5, & \text{otherwise} \end{cases} \tag{3}$$

The first key idea in SRDQN is that the capacity C of the experience replay buffer is initialized much smaller than the entire batch created from the dataset. While the IEMOCAP dataset has 12 hours of recordings, the replay buffer $\mathcal{D}$ stores 2 minutes of past data. The optimum buffer size was selected after a series of offline training at various replay buffer sizes ranging from 30 seconds to 10 minutes. The buffer size of two minutes resulted in the lowest Bellman residual values, hence it was selected. In each update step, samples from the next m time steps are pushed into $\mathcal{D}$ and an update is performed with a minibatch of m samples. Using replay buffer in this manner results in the sampling distribution μ to change gradually between gradient updates. This resembles DQN where the sampling distribution changes gradually as new online samples are pushed into $\mathcal{D}$. On the other hand, NFQ randomizes samples over the entire batch. The actors in our human-human interaction dataset display a range of emotions, resulting in the underlying policies in different scenes to vary considerably. Randomly sampling from the entire batch will result in large differences in the updates. Target updates from slow varying sampling distribution using short-term replay buffers stabilizes the training process.

The second key idea in SRDQN is the use of trajectory $(s_1, a_1, r_1, \ldots, s_m, a_m, r_m, s_{m+1})$ of length m to make an update. This is contrary to DQN where an update is made from random decorrelated samples. The replay buffer $\mathcal{D}$ in SRDQN stores samples sequentially. In each update step, a memory index i is randomly chosen from $\mathcal{D}$ (hence *random* of SRDQN). Starting from index i, m sequential samples are selected to create the minibatch (hence *sequential* of SRDQN). Thus, each update is performed from sequentially correlated samples. The problem of HRI is inherently a partially observable problem, where only observations (o) are available with some state distribution. Introducing sequentiality into the learning allows the Q-network to better estimate the underlying system state, narrowing the gap between $Q(o, a|\theta)$ and $Q(s, a|\theta)$. The importance of sequential information in partially observable MDPs has been demonstrated in the work by Hausknecht *et al.* on deep recurrent reinforcement learning [76]. Algorithm 1 summarizes the learning steps.

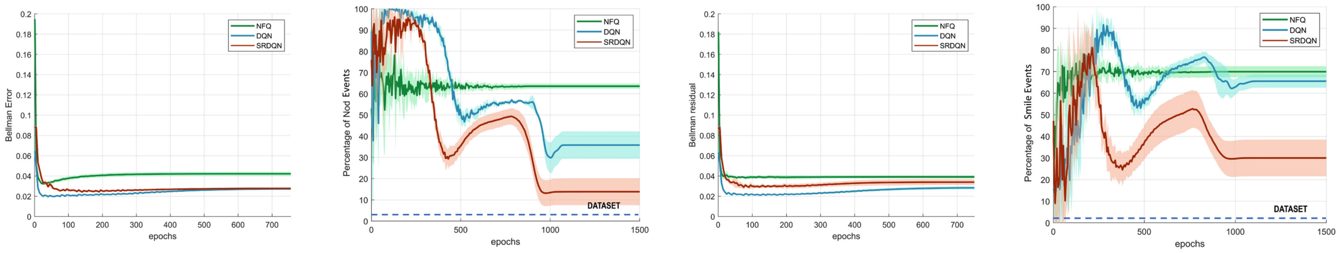
Algorithm 1. Sequential Random Deep Q-Network (Refer to Table 1 for Notations)

```
Initialize replay memory $\mathcal{D}$ to capacity $C$
Initialize $Q_\theta$ with random weights $\theta$
Initialize $\hat{Q}_{\theta'}$ with weights $\theta' = \theta$
for epoch $= 1$ to $T$ do
  while (transitions are available in the batch) do
    Get the trajectory of samples from next $m$ time steps.
    Push the $m$ samples to $\mathcal{D}$.
    Randomly choose a memory index $i$ of $\mathcal{D}$.
    Select sequential $m$ samples starting from index $i$ .
    Set $y_t = r_t + \gamma \max_a \hat{Q}(s_{t+1}, a; \theta')$.
    Determine the smooth L1 loss $\mathcal{L}$ using Equation (3)
    Perform a gradient descent step with respect to $\theta$.
    Every $N$ steps reset $\hat{Q} = Q$.
  end while
end for
```

4 Training and Evaluation

4.1 Offline Training

A multi-layer perceptron (MLP) is used to model the function approximation network. Identical architectures are used for the Q-networks of nod and smile behaviors. The function approximator takes a 209-dimensional input vector



(a) Nods: Bellman residual (b) Nods: percentage of backchannel (c) Smiles: Bellman residual (d) Smiles: percentage of backchannel

Fig. 2. Bellman residual and percentage of backchannels over the test sets at every epoch of the training.

(state feature size), has two hidden layers with 100 and 25 neurons respectively, and outputs two Q-values for the actions, ’do nothing’ and ’generate backchannel’ (nod or smile). ReLU activation function is applied between each linear layer. The Adam optimizer is used with an exponentially decaying learning rate, to update the network weights during training. We introduce the hyperparameter *reward factor* (*κ*) for scaling rewards during training. For $\kappa \leq 1$, the rewards of the batch data are modified such that

$$r'_t = \begin{cases} r_t, & \text{if } a_t = ''\text{do nothing}'' \\ \kappa r_t, & \text{if } a_t = ''\text{generate backchannel}'' \end{cases} \tag{4}$$

Our experiments show that with $\kappa = 1$, the learned policy π is dominated with the action ”generate backchannel” when evaluated with the states available in the dataset. On the other hand, the behavior policy π_b, i.e., the human policy in the dataset scarcely generates a backchannel (2.5% of actions are nods and 1.3% are smiles). We address this large deviation of learned policy from the behavior policy by penalizing the rewards. Various works have used measures like KL-divergence to determine reward penalty during learning [66], [67]. We use the simple technique of scaling down the rewards near smiles and nods, as given in Equation (4).The value of reward factor κ was selected after a series of experiments such that backchannels were reduced in number, yet without becoming fewer than those in the dataset. In all our experiments, κ is set to 0.9. The hyperparameters of the training process are given in Table 1.

Three batch reinforcement learning methods are implemented to train models for nod/smile behavior: neural fitted Q-learning (NFQ) [26], deep Q-network (DQN) [27], and Sequential Random Deep Q-Network (SRDQN). The dataset is partitioned in the ratio 4:1 for train and evaluation sets respectively. We perform the training five times for each method for 1500 epochs to address the variation seen in the converged policy when training is repeated. The greedy action from each Q-network is defined as $a = \text{argmax}_a Q(s, a)$. The final backchannel policy is determined from the votes of each of the five greedy-policies. A majority vote of 3 or above triggers a backchannel.

4.2 Offline Evaluation

Prior to deploying a policy in a human-robot interaction experiment, it is desirable to first understand the policy’s effectiveness using offline evaluation metrics. We use four different types of evaluation metrics: Bellman residual, backchannel frequency, off-policy policy evaluation and similarity to human behavior.

4.2.1 Bellman Residual

The Bellman residual [77] is calculated as the mean loss of Equation (3) over the batch of trajectories $\mathcal{B}$. A Q-function which satisfies the Bellman optimality equation, i.e., for which the Bellman error is zero for all state-action pairs, is guaranteed to be optimal, and the optimal policy is extracted by taking the action with the highest Q-value at a given state. In batch-RL, due to limited samples and Q-function approximation with neural networks, the Bellman error can only approach zero. A smaller residual implies that the learned policy is closer to the optimal policy and represents a true Q-function since it follows the Bellman equation more closely. The Bellman residual is computed for the test set at every epoch of the training. Figs. 2a and 2c show the Bellman residuals plotted against the epoch number for nods and smiles respectively. The average value of the 5 training sessions is represented by the solid line. Each curve shows a decreasing trend in the error, however, the final error for NFQ stabilizes at a value greater than those of DQN and SRDQN for both nods and smiles. The results of this first metric indicates an improved performance for DQN and SRDQN.

4.2.2 Backchannel Frequency

We also consider the fraction of states of the dataset where the new policy triggers a backchannel. Although the RL training does not aim to explicitly mimic the human in the dataset, the frequency of nods and smiles by the trained agent are expected to be comparable. The NFQ policy predicts a high occurrence of nods (63%) and smiles (70%) at the end of the training. Figs. 2b and 2d show the ratio of backchannel event prediction as the training progresses and highlights the clear disadvantage of the NFQ algorithm. Although DQN improves the nod policy, it produces comparative results for smiles. The SRDQN clearly surpasses both algorithms and predicts the nods for 13% of the states and smiles for 30% of the states. These values are closer to the actual percentage of backchannels in the dataset (1.3% smiles and 2.5% nods). Since the behaviors in the dataset do not try to explicitly maximize engagement, a higher ratio of backchannels are expected from the learned RL-policy.

TABLE 2
Off-Policy Policy Evaluation with Step-Wise Weighted Importance Sampling

	Estimated V^{π}					
	Dataset	NFQ	DQN	SL	Mirror	SRDQN
Nods	24.6	24.8	24.2	22.0	26.4	**29.8**
Smiles	24.6	26.5	20.0	21.2	26.9	**27.8**

TABLE 3
Statistics of Smile & Nod Duration (Sec)

	Nods			Smiles		
	Min	Max	Mean	Min	Max	Mean
Dataset	0.22	7.1	1.20	0.08	5.6	0.92
SL	0.10	1.42	0.42	0.1	4.97	0.55
NFQ	0.10	140	**1.50**	0.10	120	2.40
DQN	0.10	21	0.84	0.10	160	2.01
SRDQN	0.10	**6.3**	0.33	0.10	**13**	**0.77**

4.2.3 *Off-Policy Policy Evaluation*

Off-policy policy evaluation (OPE) is a class of methods to estimate the value of a policy using trajectories collected from one or more independent behavior policies [78]. We use the step-wise weighted importance sampling (step-wise WIS) method to evaluate the learned policies. It handles the distribution mismatch between the behavior policy(ies) and the evaluation policy with the importance sampling weight ρ [79]. The weight ρ_t^i at time t for the trajectory τ_i in the existing historical data is defined between the evaluation policy π_e and the behavior policy π_b as

$$\rho_t^i = \prod_{t'=0}^{t} \frac{\pi_e(a_{t'}^i | s_{t'}^i)}{\pi_b(a_{t'}^i | s_{t'}^i)}. \tag{5}$$

For N trajectories available in the batch, each with horizon T, the normalization factor at time t is defined as $w_t = \frac{1}{N}\sum_{i=1}^{N} \rho_t^i$. The value of the evaluation policy under step-wise weighted importance sampling is then given as

$$\hat{V}(\pi_e) = \sum_{i=1}^{N} \sum_{t=0}^{T-1} \gamma^t \frac{\rho_t^i}{w_t} r_t. \tag{6}$$

After training, the values of the policies resulting from NFQ, DQN and SRDQN are evaluated with the step-wise WIS estimator in Equation (6). Keeping in mind the numerical limitations, we calculate the step-wise WIS estimates over trajectory lengths of 250 samples. The WIS technique evaluates a stochastic policy whereas the greedy policy resulted from the SRDQN algorithm does not provide action probabilities. The greedy policy is converted to a stochastic policy by assigning 0.9 probability to “generate backchannel” action if the greedy policy suggests it, otherwise 0.1 probability is assigned to this action. Following the discussion of the work in [80], the behavior policy π_b is estimated using approximate nearest neighbor [81]. The off-policy policy evaluation results obtained from step-wise weighted importance sampling are shown in Table 2. The value 24.6 for the 'Dataset' corresponds to the mean sum of discounted engagement values (i.e., rewards) seen in the dataset. This value is determined independent of the actions. While the value of the NFQ policy is estimated greater than that of the dataset, DQN policy appears to perform worse, more notably with smiles. The SRDQN policy, however, is expected to accumulate more rewards, compared to the dataset and the other two policies. With a difference of approximately 5 engagement units, SRDQN policy is expected to elicit 3 more connection events per minute. We further extend the OPE analysis with two other commonly used baseline strategies, supervised learning (SL) and mirroring. When the agent is trained with a supervised learning method [43], the estimated value V^{π} of the SL-policy is found to be 22.0 for nods and 21.2 for smiles. When a mirroring policy is evaluated, that mimics speaker’s smiles and nods [40], the OPE step-wise WIS method returns a value of 26.4 for nods and 26.9 for smiles. Hence SRDQN clearly also outperforms these two baselines in terms of OPE.

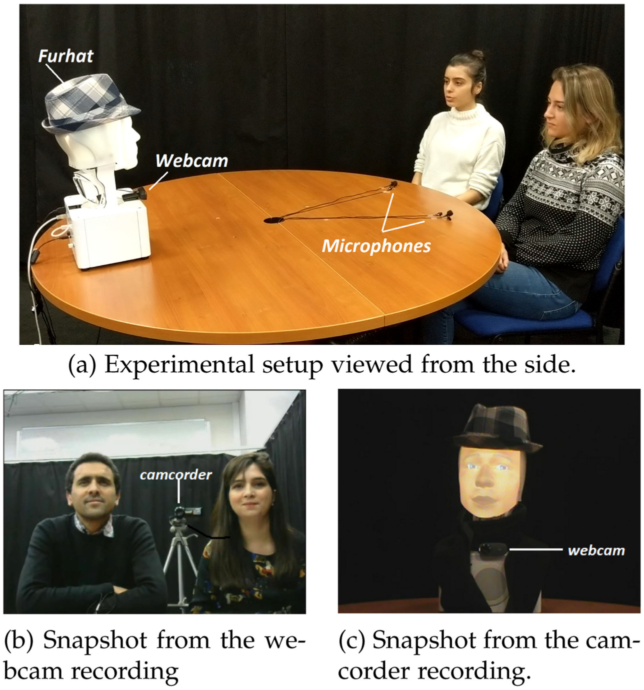
4.2.4 *Similarity to Human Behavior*

The final metric we use is the similarity of backchannel behaviour to humans in terms of duration of the backchannel. For example, a policy that results in a nod which lasts several minutes is unnatural and will result in discomfort of the user. We analyze the similarity by observing the statistical metrics of min, max and mean. Table 3 shows the statistical similarity of backchannel event duration generated by nod and smile policies to that of a human. For the nods, even though the mean nod duration for the NFQ policy matches the closest with the human nods, it also records the largest difference in the maximum duration. The maximum length of nod is noted as 140 seconds, which is socially unacceptable. The SRDQN policy displays the max duration closest to the human policy for nods. Note that, the minimum duration is bounded below in our definition of the RL formulation. We observe even poorer result for smiles with NFQ and DQN policies, with very large chunks of sequential smiles. The SL policy, on the contrary, takes fewer nod and smile decisions when compared to the dataset, giving smaller max and mean values for nod/smile duration. The statistics for mirroring policy is identical to the dataset policy, so it is not included here. Overall, it is observed that SRDQN outperforms in both measures of max and mean.

5 User Study

We designed a human-robot interaction (HRI) experiment to assess the effectiveness of the learned backchannel policy in keeping the participants engaged. Since the backchannel policy is implemented in the listening turn of the robot, the interaction is designed to motivate the participants to speak. The interaction is built on a story-shaping game where a story unfolds according to the path selected by the player. We aim to evaluate the interaction by noting the engagement indicating signals triggered during the interaction (Section 3.2), and thorough a feedback questionnaire filled by the participants.

The experiment is designed with the back-projected robot head Furhat [82], [83]. The hierarchical state machine implementation of Furhat facilitates the design of a conversation. A backchannel generation policy triggers smile and nod gestures through Furhat's event handlers. The practical implementation of nods and smiles is limited by the capabilities of Furhat. The time taken by Furhat to execute one nod is 0.5 seconds (the motion pattern of neutral-down-up-neutral). Therefore, the nod events shorter than 0.5 seconds are triggered once in our implementation. We define a 'cool down' time of 2 seconds to be used following the execution of each backchannel, during which no action is taken by the robot. The same implementation method is practiced with smiles, but with a single smile event taking 2 seconds to complete with smooth gradual changes in the lip curvature. Two models for backchannel behavior are implemented: baseline rule-based policy and test RL policy. The rule-based policy is defined such that the backchannels (nods and smiles) are generated at fixed time intervals. The test RL policy is selected to be the SRDQN trained policy which shows the best performance with the offline evaluation metrics.



(a) Experimental setup viewed from the side.

(b) Snapshot from the webcam recording

(c) Snapshot from the camcorder recording.

Fig. 3. Human-robot interaction experiment with the robot Furhat and two participants.

5.1 Experimental Design

5.1.1 Setup

The experiment is designed as a two-person game with the robot leading the interaction. The language of communication is English. During the game, Furhat addresses one participant at a time. Since the players only respond to the robot and do not talk to each other, the experience is like a one-on-one interaction similar to the training dataset. The design of the interaction with a pair of participants was preferred because sharing the experience with another user can help to increase the comfort of the participants. Moreover, an inactive participant may be motivated by his/her active partner in getting more involved in the experiment. Some snapshots of the experiment with the participants are shown in Fig. 3. In this setup, the two participants sit on fixed chairs next to each other and face Furhat around a circular table. A webcam is set in front of the participants to capture their video and a camcorder is placed behind them to record the robot's behavior. Two separate microphones are used for audio recordings of each player.

5.1.2 Procedure

A total of 26 participants took part in this user study, as pairs forming 13 groups. However, one pair faced technical issues during the experiment (the details will be given in Section 5.2.2) and hence that experiment was discarded. Among the remaining 24 participants, there were 12 females and 12 males, with ages ranging from 20 years to 40 years and mean age 27 ± 5 years. An amount equivalent to $7 was given to each participant as reimbursement for the study. When a group was called to take part in the experiment, the participants were first briefed with a short presentation. The presentation described the robot, the concept of a story-shaping game and that they would be playing two games. They were also briefed about the purpose of the experiment which was to assess which interaction with the robot was more engaging. The participants were blind to the backchannel policies implemented in the robot.

Each group was called to play the game twice consecutively on the same day, where each game differed in the generation policy of backchannels. In the baseline game (game-RB), a rule-based policy was implemented to generate backchannels in the listening mode: After a random wait period in range (0,2] seconds, either a smile or a nod was triggered with equal probability. The two backchannels were then alternated every 2 seconds till the end of the listening turn. Since the responses had a mean time length of 4.8 seconds, the average number of backchannels per turn was around 2. The backchannel generated at the end of the user's turn served as an end-of-turn feedback signal to the user. In the test game with RL agent (game-RL), the nod RL-policy and smile RL-policy were executed independently during the listening turns of Furhat. Each policy was learned with SRDQN and consisted of an ensemble of five Q-networks. As explained in Section 4.1, the final decision for backchannel was made using the majority vote from the greedy decisions of the Q-networks. Half of the groups received the game-RB as their first game and the other half received the game-RL first.

After the initial briefing, the participants were left alone in the room to run through the game. At the beginning of the interaction, Furhat introduced itself and displayed some of its capabilities such as winking and nodding. It then asked the question 'Do you think I am a well-designed robot? What do you think of me?'. This initial interaction was not part of the experiment and aimed to make the participants more comfortable and prepare them for the experiment. After the initial warm-up, the story-shaping experiment started. A video recording of the experiment

TABLE 4
Sample Dialog Between Furhat and the Players

Furhat	[Looks at Player1] Sailor, should we explore around the beach or should we rather go and discover the forest.
Player1	We should go the forest. It is better.
Furhat	Somebody seems like a nature lover. [Looks at Player2] Captain, what do you say?
Player2	I agree with Sailor. Let's discover the forest.
Furhat	It might be pretty scary in the forest. Why did you choose to go there?
Player2	Well it might have some food and fresh water.
Furhat	Looks like you both agree. [Looks at Player1]. Now we are in the forest. Sailor, what do you hear and see?
Player1	Umm. I hear a lot of birds chirping and there are many trees. And some squirrels are seen on the trees. Oh hey, there is a monkey.
Furhat	Wow. Guys, I hear some animal noises that may lead to meat. [Looks at Player2] But look in the other direction I see some fruit trees across a river. What do you think? Shall we go towards the river or hunt for animals?
Player2	Let's go towards the river. We don't have the tools to kill or capture animals.

was saved for post-experiment feedback [2]. After the experiments, the participants were asked to watch the recordings and provide their feedback by filling a questionnaire. The average time spent by participants in the first round was 7.08 minutes, and the second round was 4.85 minutes. The first round was longer because a warm-up session was conducted before the experiment.

5.1.3 Story Shaping Game

The story-shaping game is a survival adventure sketched on a desert island where the captain and a sailor are stuck after a shipwreck. The plot is designed using Furhat's dialog flow structure as a binary tree, with each node representing a scene from the story with two possible paths. Concealed behind a partition, a single trained staff controls the story transition based on the participants' responses using a Wizard of Oz interface. To enrich the interaction, along with decision questions, Furhat also asks discussion questions (e.g., the rationale behind their decisions, how they feel). In the instances when Furhat is in listening mode, the backchannel policy takes the decision of generating a nod or smile. The game tree is 5 levels deep, with half of the leaves resulting in surviving the island. An example dialogue between the participants and Furhat is shown in Table 4.

5.1.4 Architecture

The implementation of the RL backchannel policy during the experiment is shown in Fig. 4. The input data stream is the audio captured from the microphone of the participant to whom Furhat is currently attending. The audio signal is first processed to extract audio features at a rate of 40 Hz (Section 3.1). A buffer stores the audio features from past one second of data. The final state vector is created from the statistical measures as described in Section 3.1. The frequency for the backchannel decision is selected to be 4Hz. At this rate, the nod and smile policies output an action decision during Furhat's listening mode based on the state vector. Additionally, Furhat receives the audio stream along with the webcam's video stream, for its key functions such as multiple user tracking, head pose detection, automatic speech recognition etc. The audio streams from both participants and video streams from the webcam and camcorder are logged for post-experiment analysis. To measure user engagement, the connection events 'mutual gaze', 'adjacency pair' and 'backchannels' are used (refer to Section 3.2). The backchannels (laughs, smiles and nods) are annotated offline after the experiment from the video recordings and cross-checked by a second annotator. The OpenFace software [84] is used to record the mutual gaze events in real-time and the adjacency pair CE is determined online using speech events from the Furhat platform. Finally, user engagement values are calculated at 4 Hz after all the CEs are either logged online or annotated offline.

5.2 Experimental Outcome

5.2.1 Measures

The evaluation of the human-robot interaction session is performed with subject-oriented (i.e qualitative) measures and object-oriented (i.e quantitative) measures. A user feedback questionnaire is used to analyse the experimental outcome qualitatively. After the players complete the two games (game-RB & game-RL), they are asked to watch their own recordings and provide feedback with a questionnaire. The convenience of revising the interaction experience later, allows the participants to interact with the robot without any distractions during the experiment. The questions are shown in Table 5 and have similarity with those used by Sidner *et al.* in their user study [85]. They are divided into two categories: *effectiveness of the policy* and *reliability of the experiment*. The participants score each question on a 5-point Likert scale in the following ascending order: strongly disagree, disagree, neutral, agree and strongly agree. The questions in the category 'reliability of the experiment' are used to discard the sessions which are unreliable due to failure of a component of the experiment. These problems may arise due to issues in speech detection, Furhat's head motion/ facial expression capabilities or user understanding of the game. The results shown here are for the subjects who score the questions 6 to 9 with a rating greater or equal to 4. The questions in the category 'effectiveness of the policy' are designed to assess the effectiveness of the two policies in terms of engagement and naturalness.

For the object-oriented analysis, we determine the engagement values of each participant using the annotated laughs, smiles, nods, and the online logged values of adjacency pair

[2] Clips from the HRI experiment: https://mvgl.ku.edu.tr/demo-2020-backchannel-generation-with-batch-rl/

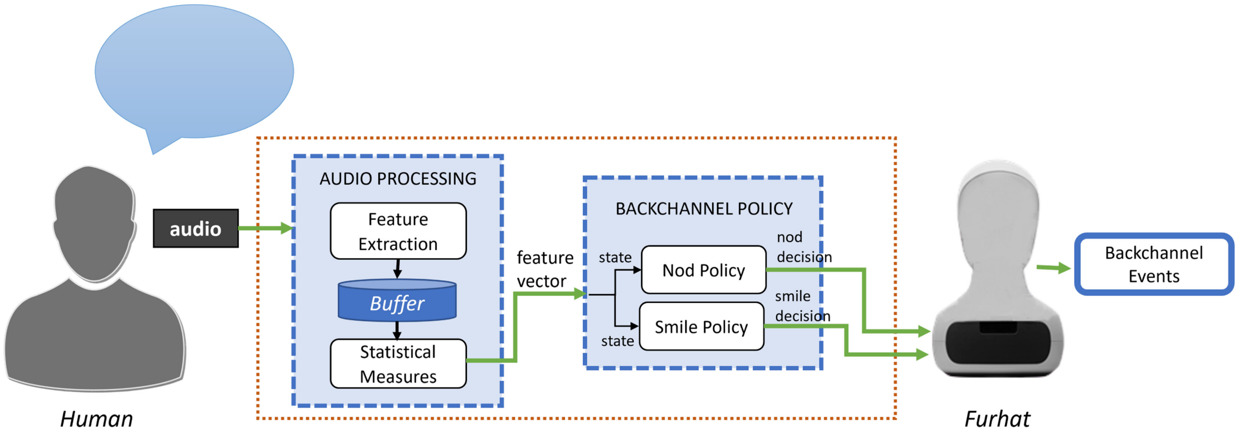


Fig. 4. Backchannel policy implementation during robot's listening turn in the human-robot interaction experiment.

and mutual gaze events. The values are recorded at a rate of 4Hz. Therefore, with each interaction lasting about 5 minutes, there are on average 1200 engagement values per interaction for each player. T-test analysis is performed to determine the significance of the results.

5.2.2 Results

As mentioned earlier in Section 5.1.2, one interaction was discarded based on the 'reliability of experiment' questions. The failure was reported due to a technical problem which occurred in Furhat's head motion during the experiment. It caused confusion in which of the two participants the robot was addressing. Besides this single discarded experiment, 12 groups (24 participants) participated in the experiment and successfully completed their interactions.

User Feedback. The 24 participants provided their feedback by filling out the post-experiment questionnaires for game-RB and game-RL. Fig. 5 shows the medians and interquartile ranges in a box plot of the ratings of each of the five questions in the category 'effectiveness of the policy'. Table 6 gives the mean ratings along with their standard deviations for each question. Overall the participants agree that both games are engaging, giving high mean scores in Q1 for game-RL (4.13) and game-RB (3.96). In order to assess each type of backchannel, Q2 inquires about the time appropriateness of nods and Q3 asks the same for smiles. Game-RL again received higher mean ratings for Q2 (3.83) and Q3 (3.58), while game-RB had lesser mean scores in Q2 (3.33) and Q3 (2.75). The average ratings of the smiles generated by the RL agent improve by a greater margin relative to the nods. While Q1 inquires in general about engagement, Q4 and Q5 are designed to understand the role of backchannels in enhancing engagement. In these questions, game-RL receives better average ratings than game-RB with a greater margin relative to Q1. The participants perceive the interaction in game-RL as more natural (4.25) relative to game-RB (3.79). Also, the higher average ratings for game-RL (4.21) relative to game-RB (3.86) in Q5 shows that the players display more interest and engagement when they interact with the RL-agent.

Furthermore, the statistical paired t-test was used to determine whether the difference between the mean ratings of game-RB and game-RL was significant. The null hypothesis stated that the values from the two groups come from the same distribution. The paired t-test results for each of the five questions are shown in Table 6. Q4 and Q5 have $p < 0.05$ and hence support the significance of the RL agent policy in ensuring more naturalness and engagement during the interaction. While the results for smiles were significant (Q3), the same was not concluded about nods. As earlier, a key finding from this analysis was that nods are less commonly perceived as ill-timed and may be accepted in a wider range of scenarios. On the other hand, smiles need to

TABLE 5
Post-Experiment Questionnaire Measuring 'Effectiveness of the Policy' and 'Reliability of the Experiment'

	Effectiveness of the policy
Q1	The interaction was engaging.
Q2	The robot's nods were timed appropriately.
Q3	The robot's smiles were timed appropriately.
Q4	The nods and smiles increased naturalness of the interaction.
Q5	The nods and smiles increased my interest in conversing with the robot.
	Reliability of the experiment
Q6	I knew what I was doing during the game.
Q7	I understood the robot well.
Q8	The robot understood me well.
Q9	The interaction went smooth

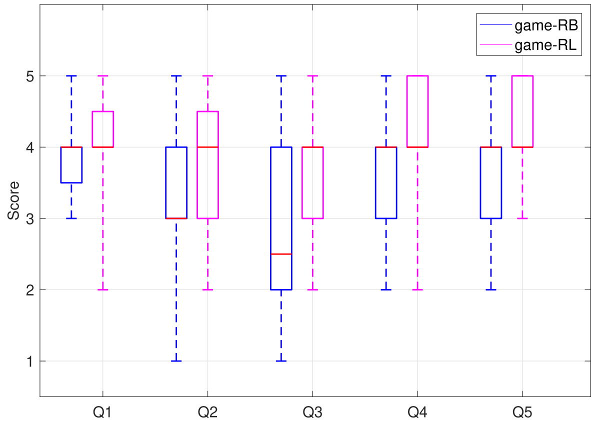


Fig. 5. Box plot for the category 'effectiveness of the policy' in the post-experiment questionnaire for the game-RB and game-RL.

TABLE 6
Statistical Analysis of Questionnaire: Mean, Standard Deviation & T-Value, P-Value of T-Test

	Game-RL		Game-RB		t-test	
	M	SD	M	SD	t(23)	p-value
Q1	4.13	0.68	3.96	0.69	1.16	0.250
Q2	3.83	0.92	3.33	1.05	1.86	0.076
Q3	3.58	0.97	2.75	1.07	3.39	0.003
Q4	4.25	0.79	3.79	0.83	2.54	0.018
Q5	4.21	0.78	3.88	0.80	2.33	0.029

be generated with more caution because context-less smiles result in a socially unacceptable behavior.

Engagement Metric. The engagement values for the two types of games (game-RB and game-RL) were compared to get an objective assessment of our method's performance. We defined a running-mean engagement $E(t)$ to observe long-term accumulation at time t as

$$E(t) = \frac{\eta_t}{t}. \tag{7}$$

where η_t is the number of CEs until time t in a given session. Fig. 6 plots the running-mean engagement $E(t)$ against time for the two types of games averaged over all the sessions belonging to that type. The plot for initial 15 seconds is not shown since we defined the first value of engagement over 15 seconds. To address different interaction lengths, all interactions were truncated to the length of the shortest interaction. The difference in the two curves shows the higher engagement of the participants playing game-RL.

We also defined a *session engagement* as a single scalar to represent engagement in each interaction, represented by $E(T)$ where T is the session length. Each of the 48 interactions was represented with a single session engagement value, where 24 samples belong to game-RB condition (M = 0.080, SD = 0.020) and 24 samples to game-RL (M = 0.089, SD = 0.022). A paired t-test was performed and indicated that game-RL sessions had significantly higher engagement values, $t(23) = 2.19, p = 0.04$.

Lastly, in order to highlight the difference between the two types of interaction, we noted the total number of smiles and nods performed by the participants during the interactions. This was manually annotated on the videos of the interactions. It was observed that total 241 smiles were generated by the users for the RL based policy versus 201 for the rule-based policy. In the case of head nods, these numbers were almost equal for the two policies.

5.2.3 Discussion

In the design of our experiment, the selection of baseline policy was an important factor. Our choice was based on the strengths and limitations of various previous techniques available in the literature. The rule-based policy employed in our user study, which is independent of user state, may look simple and repetitive. Yet, alternating between smiles and nods during interactions with an average of two backchannels per speaking turn helped alleviating the repetitive nature of the policy. During our trial experiments, the users (blind to the policies) reported that they were unable to tell if there was a policy that triggered events at regular intervals. Another common rule-based policy found in the literature is generation of a backchannel at the end of each turn. This technique is inherently part of our rule-based policy. Other baselines like 'mirroring policy' and 'supervised learned policy' are more intelligent behavioral strategies. However, our rule-based method offers some strengths over these baselines. For example, in 'mirroring policy' a backchannel is generated by the robot only after a human generates a backchannel. This puts limitation on the user to be the initiator of all backchannels. In cases where user is disengaged and barely backchanneling, the robot too will not produce any backchannels. In our baseline, it is ensured that backchannels are generated (on average twice a turn), irrespective of user's state. Similarly, as discussed earlier in the paper, since the IEMOCAP dataset is not explicitly designed to maximize engagement, a supervised learned policy will be a weak reference policy. Thus, the rule-based policy was selected as the baseline which was simple to implement and was not reported as unnatural or robot-like.

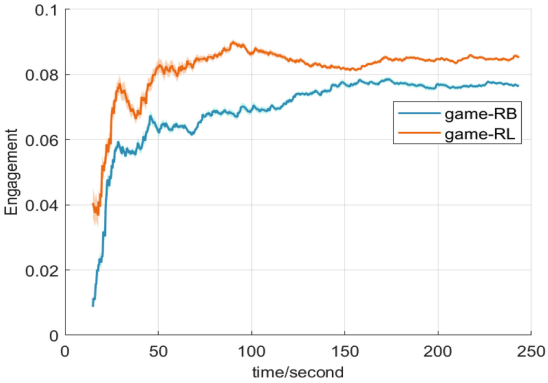


Fig. 6. Running-mean engagement $E(t)$ averaged over all sessions.

An analysis on the logs of the user study shows that Furhat remains in listening mode for approximately 30.83 minutes over the 12 experiments for testing RL-learned policies. With the RL-smile policy, a total of 138 smiles were triggered, giving an average of 4.5 events per minute. The time gap between two consecutive smile events had min = 2 seconds, max = 11.7 seconds and mean = 4.53 seconds. With the RL-nod policy, a total of 189 nod decisions were made, making an average of 6.1 events per minute. The time gap between two consecutive nods had min = 2 seconds, max = 14.5 seconds and mean = 3.6 seconds. These statistics show that while a cool-down time of 2 seconds influences the minimum gap between identical events, the average gap is still dominated by the RL policy. It is noteworthy here that while the cool-down time imposes restrictions between two smiles or two nods, there is no limitation on time gap between a smile and a nod.

The feedback questionnaire presented after the experiments provides us with an insight on how the players perceived the robot's behavior. Since participants filled the questionnaire after watching the videos of their interaction, their opinions about the robot can be thought of as a recollection of the experience. Some important conclusions may still be drawn from the feedback. The ratings from Q1 show that for both games the participants felt the interaction was engaging. Even though the t-test gave $p > 0.05$ for the ratings of Q1, the minor non-significance is interesting and encourages further research. An interesting observation here is the greater tolerance of a simple repetitive policy for nods relative to a similar policy for smiles. In the rule-based agent, the smiles score a mean value below the 'neutral' response mark. This indicates the higher complexity in the design of smile behaviors. Correspondingly, the ratings of the smiles generated by the RL agent improve by a greater margin. On the other hand, the engagement values calculated from the detection of connection events generated by the players provide us with an objective estimation of how engaged the users were during the game. Since each player experienced both games, and a significant difference was observed in the engagement values for the two types of games, we can conclude that the game with the RL policy was perceived as more engaging.

6 Conclusion

We have proposed a generation model for non-verbal backchannel behaviors of a robot, smiles and nods, to engage humans during HRI. We have demonstrated the use of recorded human-human interaction data to learn near-optimal policies with batch-RL algorithms. The results of our user study provide evidence that socially acceptable backchannel policies can be trained using only offline data. The learned policy may also serve as an initial policy for further online robot learning with humans, and hence the deployment of a possibly bothersome random policy can be avoided. The user study also highlights that different backchannel behaviors, with a similar generation pattern, are received with different levels of acceptance by users. While the policy for backchannels like nods is more flexible, more sensitive behaviors such as smiles demand a more vigorous training process. Furthermore, a key issue faced while training with batch-RL for backchannels, is the occurrence of counterfactual queries which may result in policies favoring the actions not available in the dataset, hence the distributional shift problem. To avoid this problem, we propose SRDQN with constraint as one possible solution for learning a policy that remains close to the dataset policy. Yet, more viable solutions that combine constraints on policy learning and supervision with reinforcement learning need to be explored to address the distributional shift problem as future work.

Our offline formulation for robot backchanneling is a generic framework for any robot behavior. Hence, this research work can be extended to train a variety of social behaviors such as non-verbal behaviors including gaze control, facial expressions and body gestures, and verbal backchannels like 'hmms' and 'yeahs'. We need not restrict ourselves to the goal of engagement, but multiple design objectives can be aimed. Steinfeld *et al.* list the common metrics used in the design of human-robot interaction, such as persuasiveness, trust, and compliance [86]. With accurate quantification of these metrics, they may be incorporated as the reward function into the reinforcement learning formulation.

We have demonstrated the strength of audio-based state representation for policy decisions. The state definition may be enriched with more complex user signals such as facial expressions, emotions and verbal content of speech so that partial observability of the HRI process will be alleviated. We should however note that the size of the solution space of an RL problem grows exponentially with each additional feature describing the state [87]. Hence the challenge of "the curse of dimensionality" is inevitable. Larger collections of offline trajectories, along with RL methods that identify more significant regions of state space, can be used to solve for MDPs with large solution spaces.

References

[1] K. Darling, "Extending legal rights to social robots," in *Proc. We Robot Conf., Univ. Miami*, 2012, pp. 1–25.

[2] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.*, vol. 42, no. 3/4, pp. 143–166, 2003.

[3] T. Komatsubara, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Can a social robot help children's understanding of science in classrooms?," in *Proc. Int. Conf. Human-Agent Interact.*, 2014, pp. 83–90.

[4] F. Jimenez, T. Yoshikawa, T. Furuhashi, and M. Kanoh, "An emotional expression model for educational-support robots," *J. Artif. Intell. Soft Comput. Res.*, vol. 5, no. 1, pp. 51–57, 2015.

[5] H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *J. Amer. Soc. Informat. Sci. Technol.*, vol. 59, no. 6, pp. 938–955, 2008.

[6] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *Int. J. Soc. Robot.*, vol. 7, no. 4, pp. 465–478, 2015.

[7] K. Doherty and G. Doherty, "Engagement in HCI: Conception, theory and measurement," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–39, 2018.

[8] R. Campa, "The rise of social robots: A review of the recent literature," *J. Evol. Technol.*, vol. 26, no. 1, pp. 106–113, 2016.

[9] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1019–1022.

[10] A. Hjalmarsson and C. Oertel, "Gaze direction as a back-channel inviting cue in dialogue," in *Proc. Workshop Realtime Conversational Virtual Agents*, 2012, pp. 1–8.

[11] J. J. Lee, C. Breazeal, and D. DeSteno, "Role of speaker cues in attention inference," *Front. Robot. AI*, vol. 4, 2017, Art. no. 47.

[12] C. Clavel, A. Cafaro, S. Campano, and C. Pelachaud, "Fostering user engagement in face-to-face human-agent interactions: A survey," in *Toward Robotic Socially Believable Behaving Systems*. vol. 2, Berlin, Germany: Springer, 2016, pp. 93–120.

[13] K. Lambertz, "Back-channelling: The use of yeah and mm to portray engaged listenership," *Griffith Work. Papers Pragmatics Intercultural Commun.*, vol. 4, no. 1/2, pp. 11–18, 2011.

[14] L. J. Hess and J. R. Johnston, "Acquisition of back channel listener responses to adequate messages," *Discourse Processes*, vol. 11, no. 3, pp. 319–335, 1988.

[15] L. C. Miller, R. E. Lechner, and D. Rugs, "Development of conversational responsiveness: Preschoolers' use of responsive listener cues and relevant comments," *Devlop. Psychol.*, vol. 21, no. 3, 1985, Art. no. 473.

[16] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *J. Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[17] V. H. Yngve, "On getting a word in edgewise," in *Proc. Chicago Linguistics Soc., 6th Meeting*, 1970, pp. 567–578.

[18] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *Int. J. Soc. Robot.*, vol. 6, no. 3, pp. 329–341, 2014.

[19] C. Moro, S. Lin, G. Nejat, and A. Mihailidis, "Social robots and seniors: A comparative study on the influence of dynamic social features on human–robot interaction," *Int. J. Soc. Robot.*, vol. 11, no. 1, pp. 5–24, 2019.

[20] H. Ritschel, T. Baur, and E. André, "Adapting a robot's linguistic style based on socially-aware reinforcement learning," in *Proc. IEEE 26th Int. Symp. Robot Hum. Interactive Commun.*, 2017, pp. 378–384.

[21] G. Gordon *et al.*, "Affective personalization of a social robot tutor for children's second language skills," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[22] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *J. Mach. Learn. Res.*, vol. 6, no. Apr, pp. 503–556, 2005.

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

[24] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Speech driven backchannel generation using deep Q-network for enhancing engagement in human-robot interaction," *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4445–4449.

[25] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Batch recurrent Q-learning for backchannel generation towards engaging agents," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 1–7.

[26] M. Riedmiller, "Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method," in *Proc. Eur. Conf. Mach. Learn.*, 2005, pp. 317–328.

[27] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, 2015, Art. no. 529.

[28] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Sci. Robot.*, vol. 3, no. 21, 2018, Art. no. eaat5954.

[29] O. Mubin, M. I. Ahmad, S. Kaur, W. Shi, and A. Khan, "Social robots in public spaces: A meta-review," in *Proc. Int. Conf. Soc. Robot.*, 2018, pp. 213–220.

[30] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *Int. J. Soc. Robot.*, vol. 5, no. 2, pp. 291–308, 2013.

[31] H. Robinson, B. MacDonald, and E. Broadbent, "The role of healthcare robots for older people at home: A review," *Int. J. Soc. Robot.*, vol. 6, no. 4, pp. 575–591, 2014.

[32] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *Proc. 9th Int. Conf. Rehabil. Robot.*, 2005, pp. 465–468.

[33] Y. Zhang *et al.*, "Could social robots facilitate children with autism spectrum disorders in learning distrust and deception?," *Comput. Hum. Behav.*, vol. 98, pp. 140–149, 2019.

[34] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse, "Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor," in *Proc. SIGCHI Conf. Hum. Factors ComputingComput. Syst.*, 2010, pp. 1613–1622.

[35] L. Brown, R. Kerwin, and A. M. Howard, "Applying behavioral strategies for student engagement using a robotic educational agent," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2013, pp. 4360–4365.

[36] I. de Kok and D. Heylen, "The multilis corpus–dealing with individual differences in nonverbal listening behavior," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Berlin, Germany: Springer, 2011, pp. 362–375.

[37] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Auton. Agents Multi-Agent Syst.*, vol. 20, no. 1, pp. 70–84, 2010.

[38] M. Schroder *et al.*, "Building autonomous sensitive artificial listeners," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 165–183, Apr.–Jun. 2011.

[39] K. P. Truong, R. Poppe, and D. Heylen, "A rule-based backchannel prediction model using pitch and pause information," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010.

[40] B. B. Türker, Z. Buçinca, E. Erzin, Y. Yemez, and M. Sezgin, "Analysis of engagement and user experience with a laughter responsive social robot," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 844–848.

[41] I. de Kok, D. Heylen, and L.-P. Morency, "Speaker-adaptive multimodal prediction model for listener responses," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 51–58.

[42] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," in *Proc. Int. Speech Commun. Assoc.*, 2011, pp. 2973–2976.

[43] B. B. Turker, E. Erzin, T. M. Sezgin, and Y. Yemez, "Audio-visual prediction of head-nod and turn-taking events in dyadic interactions," in *Proc. Int. Speech Commun. Assoc.*, 2018, pp. 1741–1745.

[44] J. Lee and S. C. Marsella, "Predicting speaker head nods and the effects of affective information," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 552–562, Oct. 2010.

[45] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 903–922, 2014.

[46] I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, and A. Paiva, "Modelling empathy in social robotic companions," in *Proc. Int. Conf. User Model., Adapt., Personalization*, 2011, pp. 135–147.

[47] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 1639–1645.

[48] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Robot behavior adaptation for human-robot interaction based on policy gradient reinforcement learning," *J. Robot. Soc. Jpn.*, vol. 24, no. 7, pp. 820–829, 2006.

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

[50] 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," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 154–162.

[51] L. Zhou, K. Small, O. Rokhlenko, and C. Elkan, "End-to-end offline goal-oriented dialog policy learning via policy gradient," 2017, *arXiv:1712.02838*.

[52] N. Jaques *et al.*, "Way off-policy batch deep reinforcement learning of implicit human preferences in dialog," 2019, *arXiv:1907.00456*.

[53] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. Ten Haken, and I. El Naqa, "Deep reinforcement learning for automated radiation adaptation in lung cancer," *Med. Phys.*, vol. 44, no. 12, pp. 6690–6705, 2017.

[54] X. Nie, E. Brunskill, and S. Wager, "Learning when-to-treat policies," *J. Amer. Statist. Assoc.*, vol. 116, no. 533, pp. 392–409, 2021.

[55] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electron. Imag.*, vol. 2017, no. 19, pp. 70–76, 2017.

[56] A. Kendall *et al.*, "Learning to drive in a day," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 8248–8254.

[57] A. Swaminathan *et al.*, "Off-policy evaluation for slate recommendation," in *Proc. Adv. Neural Informat. Process. Syst.*, 2017, pp. 3632–3642.

[58] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé, "Offline A/B testing for recommender systems," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 198–206.

[59] G. Kahn, P. Abbeel, and S. Levine, "BADGR: An autonomous self-supervised learning-based navigation system," 2020, *arXiv:2002.05700*.

[60] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 3406–3413.

[61] D. Kalashnikov *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proc. Conf. Robot Learn.*, 2018, pp. 651–673.

[62] S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy, "Informing sequential clinical decision-making through reinforcement learning: An empirical study," *Mach. Learn.*, vol. 84, no. 1/2, pp. 109–136, 2011.

[63] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.

[64] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2052–2062.

[65] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1179–1191, 2020.

[66] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[67] M. Geist, B. Scherrer, and O. Pietquin, "A theory of regularized markov decision processes," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2160–2169.

[68] O. Z. Bayramoğlu, E. Erzin, T. M. Sezgin, and Y. Yemez, "Engagement rewarded actor-critic with conservative Q-learning for speech-driven laughter backchannel generation," in *Proc. Int. Conf. Multimodal Interact.*, 2021, pp. 163–168.
[69] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image Vis. Comput.*, vol. 31, no. 2, pp. 137–152, 2013.
[70] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," in *Proc. 16th Int. Conf. Multimodal Interact.*, 2014, pp. 2–9.
[71] A. Schmitt, T. Polzehl, and W. Minker, "Facing reality: Simulating deployment of anger recognition in IVR systems," in *Proc. Int. Workshop Spoken Dialogue Syst. Technol.*, 2010, pp. 122–131.
[72] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, I. Poggi, and U. R. Tre, "Engagement capabilities for ECAs," in *Proc. AAMAS Workshop Creating Bonds ECAs*, 2005, pp. 1–8.
[73] R. Ishii and Y. I. Nakano, "Estimating user's conversational engagement based on gaze behaviors," in *Proc. Int. Workshop Intell. Virtual Agents*, 2008, pp. 200–207.
[74] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Proc. IEEE/ACM 5th Int. Conf. Hum.-Robot Interact.*, 2010, pp. 375–382.
[75] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.
[76] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," *CoRR*, vol. 7, no. 1, pp. 1–9, 2015.
[77] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Proc. Mach. Learn. Proc.*, 1995, pp. 30–37.
[78] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT press, 2018.
[79] C. Voloshin, H. M. Le, N. Jiang, and Y. Yue, "Empirical study of off-policy policy evaluation for reinforcement learning," 2019, *arXiv:1911.06854*.
[80] A. Raghu *et al.*, "Behaviour policy estimation in off-policy policy evaluation: Calibration matters," 2018, *arXiv:1807.01066*.
[81] V. Hyvönen *et al.*, "Fast nearest neighbor search through sparse random projections and voting," in *Proc. IEEE Int. Conf. Big Data*, 2016, pp. 881–888.
[82] S. Al Moubayed, J. Beskow, and G. Skantze, "The Furhat social companion talking head," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 747–749.
[83] S. A. Moubayed, G. Skantze, and J. Beskow, "The Furhat back-projected humanoid head–lip reading, gaze and multi-party interaction," *Int. J. Humanoid Robot.*, vol. 10, no. 01, 2013, Art. no. 1350005.
[84] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.
[85] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artif. Intell.*, vol. 166, no. 1–2, pp. 140–164, 2005.
[86] A. Steinfeld *et al.*, "Common metrics for human-robot interaction," in *Proc. 1st ACM SIGCHI/SIGART Conf. Human-Robot Interact.*, 2006, pp. 33–40.
[87] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

Nusrah Hussain received the BSc degree in electrical engineering from the University of Engineering and Technology Lahore, in 2010 and the MS degree in systems engineering from the Pakistan Institute of Engineering and Applied Sciences, Islamabad, in 2012. She is currently working toward the PhD degree with the Electrical & Electronics Engineering Department, Koç University, Istanbul, Turkey. Her research interests include human-robot interaction, social robotics, audio-visual signal processing and deep learning.

Engin Erzin (Senior Member, IEEE) received the BSc, MSc, and PhD degrees from the Bilkent University, Ankara, Turkey, in 1990, 1992 and 1995, respectively, all in electrical engineering. During 1995-1996, he was a postdoctoral fellow with Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a member of technical staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he is with the Electrical & Electronics Engineering and Computer Engineering Departments of Koç University, Istanbul, Turkey. Engin Erzin is currently a member of the *IEEE Speech and Language Processing Technical Committee* and associate editor for the *IEEE Transactions on Multimedia*, having previously served an associate editor of the *IEEE Transactions on Audio, Speech & Language Processing* (2010-2014). His research interests include speech signal processing, audio-visual signal processing, human-computer interaction and pattern recognition.

T. Metin Sezgin received the graduate summa cum laude degree with honors from Syracuse University, in 1999, the MS degree from MIT AI Lab, in 2001, and the PhD degree from MIT, in 2006. He subsequently joined the Rainbow group , University of Cambridge Computer Laboratory as a postdoctoral research associate. He is currently an associate professor with the College of Engineering, Koç University. His research interests include intelligent human-computer interfaces, and HCI applications of machine learning. His research has been supported by international and national grants including grants from DARPA (USA), and Turk Telekom. He is a recipient of the Career Award of the Scientific and Technological Research Council of Turkey.

Yücel Yemez received the BS degree from Middle East Technical University, Ankara, in 1989, and the MS and PhD degrees from Boğaziçi University, Istanbul, respectively, in 1992 and 1997, respectively, all in electrical engineering. From 1997 to 2000, he was a postdoctoral researcher with the Image and Signal Processing Department of Telecom Paris (ENST). Since September 2000, he is with the Computer Engineering Department, Koç University as a faculty member. He is currently an associate editor of *Elsevier's Graphical Models Journal*. His research interests include computer vision, human-computer interaction, machine learning and multimedia signal processing.