How Did We Miss This? A Case Study on Unintended Biases in Robot Social Behavior

Maria Teresa Parreira
KTH Royal Institute of Technology
Sweden

Sarah Gillet
KTH Royal Institute of Technology
Sweden

Katie Winkle
Uppsala University
Sweden

Iolanda Leite
KTH Royal Institute of Technology
Sweden

ABSTRACT

With societies growing more and more conscious of human social biases that are implicit in most of our interactions, the development of automated robot social behavior is failing to address these issues as more than just an afterthought. In the present work, we describe how we unintentionally implemented robot listener behavior that was biased toward the gender of the participants, while following typical design procedures in the field. In a post-hoc analysis of data collected in a between-subject user study (n=60), we find that both a rule-based and a deep learning-based listener behavior models produced a higher number of backchannels (listener feedback, through nodding or vocal utterances) if the participant identified as a male. We investigate the cause of this bias in both models and discuss the implications of our findings. Further, we provide approaches that may be taken to address the issue of algorithmic fairness, and preventative measures to avoid the development of biased social robot behavior.

CCS CONCEPTS

• Computer systems organization → Robotics; **• Human-centered computing** → *Empirical studies in collaborative and social computing*; **User studies**; **• Computing methodologies** → *Supervised learning*.

KEYWORDS

ethical HRI, AI fairness, gender bias, machine learning, non-verbal behaviors

ACM Reference Format:
Maria Teresa Parreira, Sarah Gillet, Katie Winkle, and Iolanda Leite. 2023. 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 (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3568294.3580032

1 INTRODUCTION

One goal of the Human-Robot Interaction (HRI) community is to explore how we can create social behaviors for our social robots that benefit the interaction between humans and robots. For example, the field has suggested techniques to automatically generate a robot's gaze [15, 16], empathetic behavior [40], or backchannels [21, 53]. These techniques often propose a set of rules to form a heuristic [38, 52] or learning techniques that are trained on data from human-human interactions [13, 36].

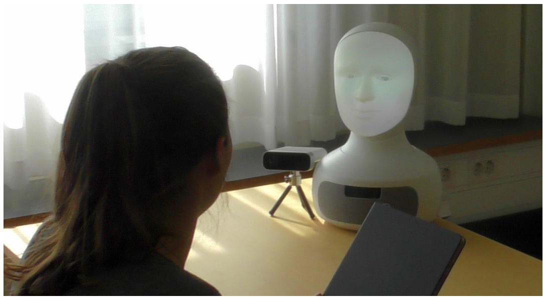

Figure 1: Overview of the experimental set-up in Parreira et al. [41]. Participants were invited to think aloud while completing tasks on the tablet. In this work, we examine the conditions in which participants directed their words to a robot. The robot was performing listening behavior in the form of backchanneling, either following a heuristic (*NaïveL*) or learned (*DataL*) behavior model.

Pioneering work by Buolamwini and Gebru [6] has shown that datasets can be a source of bias in machine learning methods. These biases reflect a variety of biases present in society, namely skin color in combination with gender [6], gender in recruiting tools [9], and race when predicting the likelihood of future crimes [1]. Recently, the HRI community started to explore the effect of biased robots and found that even seemingly obvious and objectively biased behavior goes unquestioned, with unfair outcomes instead being post-hoc rationalised in a way that reflects (gender) stereotypes [20].

Here, we discuss how pursuing a very typical, data-driven approach to the development of a robot listener behavior (production of backchannels, which can serve to indicate attentiveness) resulted in models that acted differently with participants with different gender identities. In a between-subject study [41], we explored two different methods to predict when to emit backchannels: a commonly used heuristic [53] and a deep learning model trained on human-human conversational data. On running these models online in the user study, we (authors 1 and 2) independently observed that the robot appeared to be behaving differently across female and male participants. Post-hoc analysis subsequently confirmed that

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580032

both approaches produced more backchannels when the speaker was male. It is easy to imagine observant bystanders perceiving the robot as hence being more attentive to men than to women – which would not be an isolated case, given ongoing gender disparities in digital skills development [54], inclusion within computing and robotics [50, 55] and criticism of social robots/virtual agents propagating harmful gender stereotypes [51, 54, 56].

In the present work, we describe and reflect on our system design process in order to provide insight into the potential pitfalls when developing social behaviors for robots. We analyze the robot's behavior in conjunction with key features from the dataset and our participant pools, in order to understand why this bias emerged. Finally, we contribute with a set of design preventative measures that researchers may take into account when adopting data-driven approaches to the generation of robot behavior, in order to avoid repeating our mistake and deploying unfair, biased robot social behavior.

2 BACKGROUND

The emergence of biases in social behavior models is a multidisciplinary phenomenon, which ought to be contextualized. Below, we shed light on human-human listening behavior, approaches to automatize this behavior in artificial agents, and why our findings are not an isolated issue.

2.1 Backchannels in human-human interactions

Backchannels (BCs) are short vocal or non-vocal expressions of a listener that are not meant to interrupt the turn of the current speaker [17]. They play an important conversational role, by signaling attentiveness or emotion to a speaker. Backchanneling has also been found to impact the perceived personality and rapport building [5, 23]. While BC behavior has been shown to be correlated with the personality of the listener [22], evidence indicates that the gender identities of the speaker and listener also play a role. A substantial number of studies reports that participants identifying as female backchannel more than participants identifying as male [4, 45, 46], and female-identifying speakers generally receive the most backchannels [4]. Adjacent factors such as status or dominance also influence the amount of backchanneling [33, 45]. Interestingly, Mulac et al. [35] found that female and male observers might interpret the function of backchannels in different ways. Females tended to associate backchannels with interest, whereas males interpreted them as a sign of uncertainty.

The literature described above highlights the role of BCs and indicates that the speaker's gender identity may impact the backchanneling that occurs in a conversation; also given its impact on the perception of the interaction, we should carefully observe differences in BC behavior when executed on a social robot. We further discuss these implications in Section 4.

2.2 Listener Behavior in HRI

Prior work explored multiple approaches to the development of listening behavior in robots. Rule-based methods are common and predict backchannel behavior by monitoring the speaker's prosodic features, e.g., the pitch [38, 52, 53]. However, hand-crafted heuristics can be limited, which is why other authors adopt automated learning methods [13, 27]. For example, Okato et al. [39] designed a Hidden Markov Model (HMM) based on prosodic patterns to detect when to emit a BC. Morency et al. [32] also used a HMM with multimodal (audiovisual) input features to predict BC timing.

More recently, the development of listener behavior has been making use of deep learning techniques. A common architecture is Recurrent Neural Networks (RNNs), as they capture the temporal dependencies of continuous signals (i.e., retain "memory" of previous inputs). Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) layers, in multimodal input models (acoustic features, video features, or word history) have been leveraged for the development of idle behavior in virtual agents or robots [21, 48]. In addition to implementing a listener model using LSTMs, Murray et al. [36] also suggested a method for data augmentation that positively impacts BC prediction. RNNs have also been used to inform turn-taking behavior [19, 44]. Other interesting approaches include the use of such as semi-supervised learning [26], or reinforcement learning [25, 47].

2.3 Bias and Critical Reflection in HRI and Related Fields

Buolamwini and Gebru's *Gender Shades* project [6] drew major attention to the existence of bias within facial recognition systems – a key 'building block' designed to underpin human-machine interactions. Their work pointed, in part, to non-diverse and/or biased datasets as being one major cause of biased algorithms more generally. Other notable examples include, e.g. "gender recognition" systems that seemingly identify the gender of a person based on their surroundings (kitchens and handbags being synonymous with women, skateboards and surfboards with men) [31]. Buolamwini's subsequent work with the *Algorithmic Justice League*[1] aims to create 'cultural movement towards equitable and accountable AI'.

A number of authors spanning science and technology studies, human-computer interaction and social sciences have made similar calls for change. For example, Strengers and Kennedy [51] call for a 'feminist reboot' of assistive technologies (including robots) to ensure more ethical designs which challenge, rather than propagate, harmful gender norms and stereotypes. Ruth Benjamin[3] identifies a need for 'sociologically informed skepticism' when examining new technologies, based on her study of AI-powered technologies as the *New Jim Code* – hiding and even speeding up discrimination under the veil of machine neutrality.

Perhaps driven in part by these calls, recent works within HRI have provided critical reflections on how research is conducted, aiming to make explicit any assumptions and potential biases adjacent to the robots/interactions we're developing. For example, concerning data diversity, Winkle et al. [55] recently reviewed gender diversity in HRI research participation to date, finding an over-representation of men and an under-representation of women (moderate) and non-binary persons (very large). Works on (gendered) robot design/perception have probed the ways in which gender stereotypes, combined with a potential lack of diversity in research/design teams, might be manifesting in current robot

[1] https://www.ajl.org/

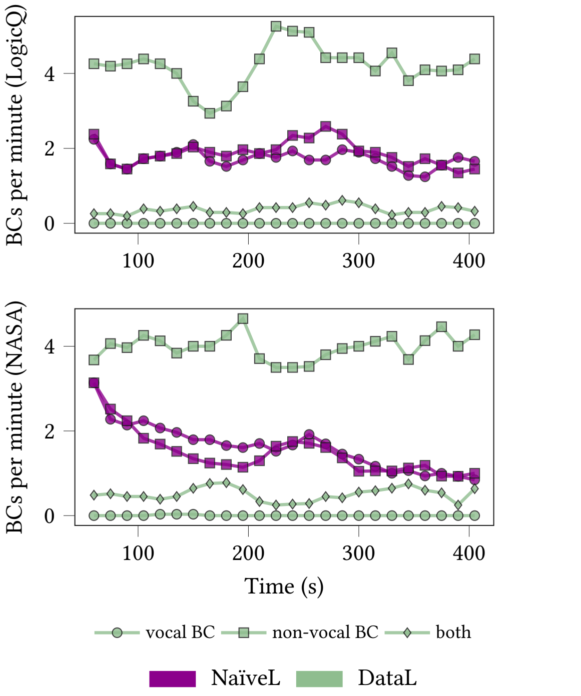

Figure 2: Backchanneling behavior over time from the data-driven model (*DataL*, in green) or heuristic model (*NaïveL*, in purple). The different types of BCs emitted are discriminated. Data shown is averaged over all participants in that condition. Top - LogicQ task; bottom - NASA task.

design and applications [14, 42, 56]. Concerning the potential impacts of biased robot behavior, work by Hitron et al. [20] found that participants failed to recognize gender bias in a robot moderating a debate, instead explaining its unfair behavior through arguments that propagate gender stereotypes. This finding particularly motivates our current work, as it demonstrates just how *easily* biased behavior can go unnoticed.

3 GENDER BIAS IN LISTENER MODELS

This paper focuses on a post-hoc realization of robot behavior biases during a user study. As such, we provide only an overview of the study and model implementation. A more detailed description is provided in Section 6.

3.1 User Study

In Parreira et al. [41], we set to explore how a social robot could improve a *rubber duck debugging* session. This concept, which originated from a story in the book The Pragmatic Programmer [24], encompasses the idea that a rubber duck can assist the process of "debugging" a program by serving as a listener when a person explains the code aloud step-by-step. We wanted to evaluate the effects of replacing the rubber duck, an unresponsive listener, with an attentive listener robot, in a Think-Aloud Problem-Solving (TAPS) [29] session.

In a between-subject study with three conditions, we evaluated how two different robot listening behaviors and an inanimate object - a rubber duck - affect the outcome and perception of two distinct think-aloud problem-solving tasks, a deductive logic quiz (LogicQ task) and an open-ended question (NASA task). The setup can be seen in Figure 1. We used a Furhat robot[2] with the William voice from CereProc[3], including its backchannel sounds.

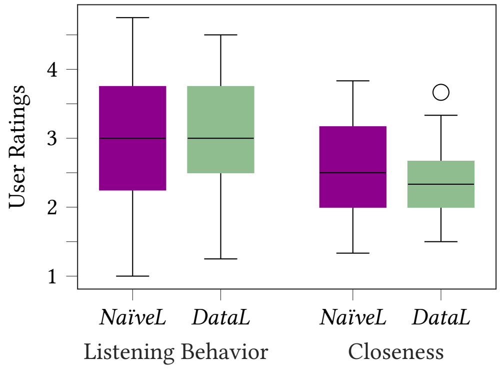

Figure 3: Participants' rating of the robot's listening behavior and closeness to the robot (from Murray et al. [36]).

3.1.1 With-robot Conditions. This study utilized two different listener behaviors. In the **Naïve Listener Robot (NaïveL)** condition, the social robot displayed listening behavior generated by handcrafted heuristics. Conversely, in the **Data-driven Listener Robot (DataL)** condition the social robot displayed listening behavior that was learned through machine learning methods using a human-human conversational corpus.

3.2 Listener Behavior Implementation

The listener behavior implemented consisted of the robot's gaze and backchanneling (vocal and/or non-vocal, in the form of nodding). In the present work, we focused solely on the analysis of the backchanneling behavior produced, whose implementation we describe below.

3.2.1 Rule-based Model. The **NaïveL** condition made use of the implementation of a well-known heuristic developed by Ward and Tsukahara [53]. This impactful work made use of corpora of English and Japanese human conversation data and suggests a set of rules for each language that determine the appropriate time to emit a backchannel. The heuristic is based on prosodic cues, i.e., pitch. The first two rules are the most influential as indicators for backchannel opportunities: (1) the pitch has to fall below the 26th percentile of the overall pitch distribution and (2) continue in this region for at least 110ms. We reproduce the full set of rules and give further implementation details in Section 6.1.

3.2.2 Data-Driven Model. In addition to a rule-based model, we also deployed a deep learning listener model which learned backchanneling behavior from a corpus of human-human conversational data. The model is split into two models: BC timing (*when to backchannel*) and BC type (*which type of BC to emit* - vocal or non-vocal (nod)). We give a brief overview of the deep learning listener model here

[2] https://furhatrobotics.com/
[3] https://cereproc.com/

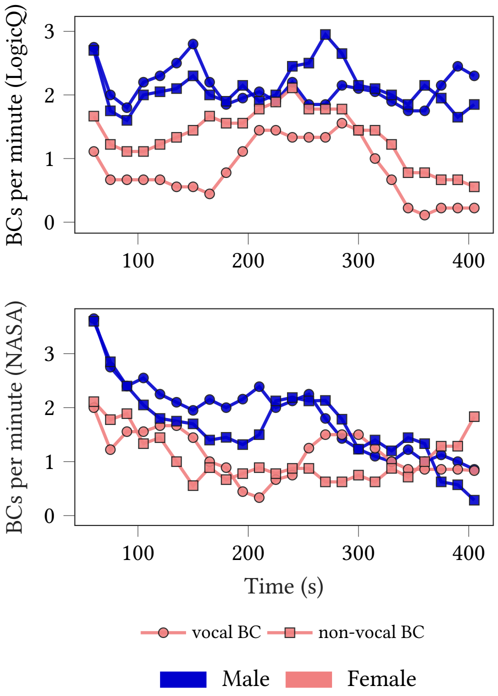

Figure 4: Backchanneling behavior over time from the rule-based model (*NaïveL*) across genders (male - blue, female - red). The different types of BCs emitted are discriminated. Data shown is averaged over all participants in that condition. Top - LogicQ task; bottom - NASA task.

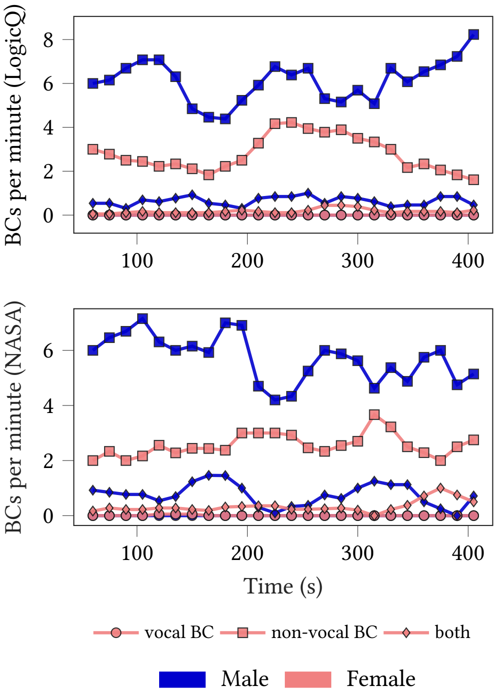

Figure 5: Backchanneling behavior over time from the data-driven (*DataL*) across genders (male - blue, female - red). The different types of BCs emitted are discriminated. Data shown is averaged over all participants in that condition. Top - LogicQ task; bottom - NASA task.

and refer the reader to Section 6.2 for a full problem and model description as well as training procedure and performance.

The model was built on a subset of the Cardiff's Conversation Database (CCDb)[2]. For training, we selected eight of 30 conversations based on the thoroughness of annotation. From these eight conversations, we extracted the states and actions for the deep learning model. The states were formed by a 34-item feature vector extracted at 2Hz from the audio data, comprising of mel-frequency cepstrum coefficients (MFCC) and prosody features. The deep learning model takes this feature vector as input to generate a backchannel action as an output. The robot could perform the same actions as those performed by the *NaïveL* robot (Sec. 6.1) - vocal utterance, nod of varying amplitude, or a simultaneous combination of both.

3.3 Observed Backchanneling Behavior

To evaluate what behavior the models generated and how it was perceived by participants, we looked into how the robot acted and how it was perceived in the *NaïveL* and *DataL* conditions. We provide the data and R markdown files used for the analysis[4].

3.3.1 Participants. In Parreira et al. [41], a total of 101 participants were recruited through posters, flyers, social media platforms, and word of mouth. Ages ranged from 19-76 years (M = 26.4, SD = 7.6). 53 participants identified as male and 48 as female, with a total of twenty-nine different nationalities. Each condition where the robot was present had the following participant demographics: $NaiveL - N = 29$ $(9F, 20M)$, ages 26.3 ± 7.4; $DataL - N = 31$ $(18F, 13M)$, ages 25.3 ± 5.3.

3.3.2 Robot Behavior. To understand how the robot acted during the interactions, we analyzed the log files from each interaction. We defined **Backchannel Frequency per minute** as the total BCs executed by the robot every 60 seconds, calculated as a sliding window with hop length of 15s. Figure 2 shows the average backchannel frequency across time for each task. The *NaïveL* robot displayed similar values for vocal and non-vocal backchanneling, but the frequency decreased as task time increased. The *DataL* robot, on the other hand, favors non-vocal backchannels (nodding), which is displayed in much higher frequency, as well as a combination of both nodding and vocal utterances.

3.3.3 User Reporting of Robot Behavior. The participants in conditions where the robot was present (*NaïveL, DataL*) were asked to evaluate the robot in the post-experiment questionnaire. They rated the robot's **Listening behavior** and **Closeness** [36], as well as **Social attributes** (from RoSAS [8]).

A one-way ANOVA showed no significant differences between conditions for both the **Closeness** and **Listening behavior** dimensions of robot behavior [36] ($F(1, 58) = 2.07, p = 0.16$ and $F(1, 58) = 0.56, p = 0.46$, respectively). Fig. 3 illustrates these findings. An analysis of the social attributes [8] also did not reveal a significant difference between the *NaiveL* and *DataL* conditions (one-way ANOVA for the **Competence** dimension, $F(1, 58) = 0.46, p = 0.50$, and Wilcoxon rank sum test for the **Warmth**,$W = 543.5, p = 0.10$ and

[4] https://github.com/mteresaparreira/gender-biased-robot-duck

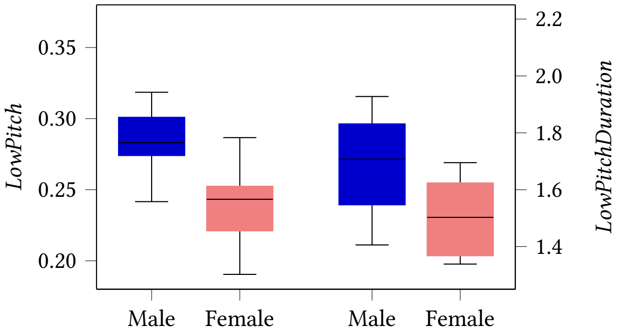

Figure 6: Low pitch cue is different across genders. Left: ratio of audio samples under the 26th percentile level. Right: average sequential time below that level (in units of computation, approximately 10 ms/unit).

Discomfort, $W = 410, p = 0.55$, dimensions, after a Shapiro–Wilk test of normality revealed these variables did not have normal distributions, $p < 0.05$).

3.3.4 Different Backchannel Frequency for Different Genders... When deploying the models online, in the user study, the robot appeared to behave dissimilarly when the participant identified as male or female. In the *NaïveL* condition, researchers running the experiment noticed the frequency of backchannels emitted appeared to increase with male participants. Indeed, when breaking Fig. 2 by gender, we can observe that, for both the rule-based model (Fig. 4) and the data-driven model (Fig. 5), the robot displayed higher levels of BC feedback.

We evaluated the effect of **Gender** on **BC Frequency** with an ANCOVA, after controlling for **Speech-to-silence ratio** (for each participant and task, time spent speaking over time spent silent).

NaïveL. Gender was a significant predictor of **BC Frequency**, $F(1, 28) = 13.37, p = 0.001$.

DataL. Gender was also a significant predictor of **BC Frequency**, $F(1, 30) = 7.58, p = 0.01$, as well as Speech-to-silence ratio, $F(1, 30) = 63.07, p = 0.008$.

3.3.5 ... But Participant Perception Did Not Change. Following the results in Section 3.3.4, we investigated if these differences in behavior reflected in differences in how users evaluated the robot. For each measure - **Listening Behavior, Closeness, Warmth, Competence, Discomfort**-, we ran an ANCOVA to evaluate the effect of **Gender**, controlling for **BC Frequency**.

For both robot behaviors, we found no significant effects of **Gender** or **BC Frequency** on any of the above-mentioned measures.

3.4 Debugging Robot Behavioral Differences

3.4.1 When Rules Don't Apply. Ward and Tsukahara [53] defined a set of rules (Section 6.1) based on three main components: participant's voice pitch, voice activation detection (VAD), and time since the last backchannel was emitted. The distribution of voice pitches is different across sexes [30]; thus we identified the calculation of the 26th-percentile pitch level as being one possible source for differing robot behaviors across gender.

We investigated this hypothesis by extracting the pitch and 26th percentile value for the two tasks of 22 participants (6F, 16M, other audio recordings were corrupted). As detailed in Section 6.1, the pitch is updated with the last 10 ms of audio. This instantaneous pitch value is compared to the 26th-percentile level (rule **P1**). We investigated the percentage of pitch values that fell below this level per participant (**LowPitch)** and examined if it differed by gender (Fig. 6, left). Additionally, as per rule **P2**, we investigated the average duration the pitch stayed below the 26th percentile (**LowPitchDuration**, Fig.6, right). The percentile is calculated over the last 50 s of audio, and thus its value is not fixed - hence why we see values for **LowPitch** which are above or below 26%.

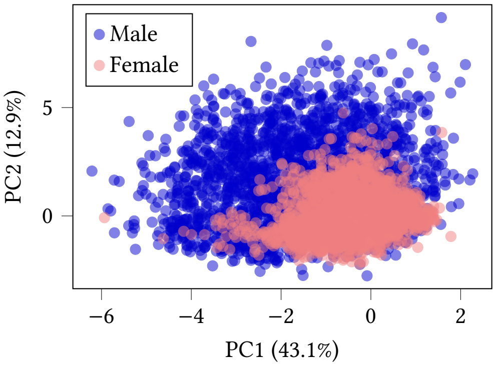

Figure 7: Principal components that explain the most variance (56%) in the state space, by gender.

For both rules, male participants trigger more positive opportunities for backchanneling (Wilcoxon rank sum test, **LowPitch** - $W = 11, p < 0.01$, **LowPitchDuration** - $W = 20, p = 0.04$, after a Shapiro–Wilk test of normality revealed these variables did not have normal distributions, $p < 0.05$).

3.4.2 State Space Also Differs. As described in Section 6.2, the data-driven model takes as input a 34-dimension prosody feature vector, extracted from the audio of the participants. In order to investigate what could be causing the differences in the *DataL* robot behavior, we extracted the input feature vectors across time for all participants in this condition (18F, 13M). To best observe the differences between genders in the state space, we ran a Principal Component Analysis (PCA), which allows us to perform dimension reduction. Fig. 7 shows the two principal components which explain 56% of variance. While there is some overlap, features from female participants appear less dispersed. A one-way ANOVA shows a main effect of **Gender** for the distribution of values from the first two principal components (**PC1**: $F = 2869.3, p < 0.001$, **PC2**: $F = 5254.9, p < 0.001$).

3.4.3 Is it all in the dataset?

NaïveL. For the extraction of the rules, Ward and Tsukahara [53] use a corpus of 68 minutes of dyadic conversations in English, with a total of 12 speakers. Gender diversity amongst the 12 speakers is low, made up of 10 male and 2 female speakers.

DataL. To train a deep learning listener model that could adjust to different user tasks, we looked for corpora of unscripted, non-topic-bounded interactions that were publicly available. We used Cardiff's Conversation Database (CCDb)[2], which has approximately 150 minutes of dyadic conversations. The thirty conversations occurred between 16 different speakers (12 M, 4 F). For training and offline testing, we selected the subset of 8 conversations annotated for facial expressions and utterances. These interactions took place among 6 speakers - all male participants.

4 DISCUSSION

In this section, we discuss our findings and leave suggestions for measures to avoid replicating the deployment of unfair robot social behavior models.

4.1 Biased Backchanneling

In Ward and Tsukahara [53], the authors provide a discussion of the rationale and limitations of the performance measures used to evaluate their rule-based model, as well as the challenge of accurately predicting backchanneling behaviors given how dynamic social valence and rapport are during an interaction. They advance hypotheses about the communicative functions of the low pitch regions as defined by rules **P1** and **P2**, namely that the speaker considers they have transmitted information. However, as per the data collected in our previous study [41], male speakers produce these cues more frequently (Fig. 6), leading to a more responsive listening behavior for men than for women. Note that speech-to-silence ratio was not deemed a predictor of the frequency of BCs emitted (Section 3.3.4), which means that the observed differences in listening behavior are not due to higher speaking engagement on the part of male participants. In spite of the author's thorough work and analyses, a lack of gender diversity within their participant poll may be one reason why this aspect was missed.

For the **DataL** condition, we note that, in spite of the principal component distributions observed, *speech-to-silence* ratio was also a predictor of the robot backchanneling behavior. This may be a good indication that this model's output was informed by other important aspects of the interaction.

Both the models deployed were biased towards the overproduction of BCs in a conversation if the speaker was a male. Interestingly, this is contrary to what can be observed in human-human conversation, as female-identifying participants tend to receive more backchannels [4]. This reveals that the bias is not a reflection of sociological biases, but rather emerges from the methods and models used in this study.

Nature and Availability of Human-Human Conversational Data. The collection of conversational data is frequently the first step to developing listening behaviors. Some interesting corpora of dyadic interactions in English include CCDb [2], IEMOCAP [7] and D64 [37], among others [49]. We were looking for unscripted, non-topic-bounded interactions that were publicly available. Not many corpora exist that fulfill these criteria, and even fewer datasets are mindful of gender balance among participants. We highlight IEMOCAP[7], which collected data from 5F, 5M actors. The literature on the development of artificial listener behavior is, however, abundant in examples of gender imbalance within the participant pool. For example, de Kok et al. [13] (3F, 29M), Poppe et al. [43] (4F, 16M), Morency et al. [32] (67F, 37M), Okato et al. [39] (0F, 22M), Murray et al. [36] (9M, 3F). Other works don't disclose gender (e.g. Kawahara et al. [27]). In order to foster robustness of robot social behavior models, more attention needs to be given to gender distribution and potentially other demographic and social factors when recruiting participants.

4.2 We Built a Biased Robot

The present work focuses on a post-hoc exploration of the causes of a bias in robot social behavior. Backchanneling behavior in a casual social interaction with a healthy, neurotypical adult population may not seem like a high-stakes setting for biases, but even within this context, *a posteriori* identification of biases has potentially devastating consequences. Pitch cues are a source of gender stereotyping [28], and women tend to recall non-verbal behavior more than men [18]. Failure to recognize this bias by participants (gender was not a predictor of how the robot was evaluated) was not surprising, as it has been observed before [20]. Rather, it identifies the risk of biased behaviors going unnoticed/unquestioned, not only by participants but also by *researchers*, who might look at such lack of difference as evidence that the robot 'worked equally well for everyone'.

Our own failure to prevent these biases came as an unpleasant surprise and made us wonder what could have been done differently. The observed differences in robot behavior are without a doubt unacceptable, but it remains unclear how the robot should behave instead. We see three contrasting approaches that can be taken to develop (hopefully) more "fair" data-driven robot social behaviors: **a)** reproducing the differences observed in human-human interactions, **b)** attempting to treat users of all gender identities exactly equally, or **c)** applying the concept of *algorithmic reparation* [10] to overcompensate for those differences observed in human-human interactions. Below, we will discuss each point separately. We would like to note that, in all cases, potential differences in perception of social behaviors as discussed in Section 2.1 need to be considered.

Reproducing observed differences In Section 2.1 we discuss observations on backchanneling behavior as perceived from and produced by male-identifying and female-identifying conversation partners. Additionally, while some works discuss that male-identifying conversation partners might use backchannel behavior to exhibit control [35], the literature remains undecided if the given experimental data allows for such a conclusion. Future work should therefore be aware of the discussion on societal biases even in seemingly small social behaviors when developing these. Further, future work needs to address if robots with different gender identities might raise different expectations in users given the differences in producing backchannels based on gender identity.

Equal treatment Future work should also explore the feasibility of equal treatment across genders. We note, however, that this implies the robot will violate expectations of gender norms through its executed behavior, since there are implicit gender biases in human-human interactions (e.g., different backchanneling behavior in males and females [4]).

Overcompensation Recently, Davis et al. [10] suggest the paradigm of *algorithmic reparation*, which calls for overcompensation in case of historical injustice. For example, given that fewer women were hired for tech sectors, more women than men must be hired in the future to overcompensate for previous injustice. Future work will need to understand to what extent social behaviors in conversations can offer a possible place to overcompensate for societal biases in these interactions.

4.2.1 Preventative Measures. With the above discussion and our findings in mind, we suggest specific design guidelines that may be applied to avoid embedding gender biases into robot social behavior:

(1) Carefully examining the dataset and considering reducing or recollecting the data to form a pool of gender-diverse participants.
(2) Testing developed behaviors online and offline whilst
  - being sensitive to pilot testers (and observing researcher) impressions of the robot's behaviors across people of different identities – here we have explored only gender, but such work requires an intersectional approach;
  - quantifying behavior by running pilot studies or performing offline evaluations on a dataset to objectively observe the robot's behavior towards participants with different (gender) identities;
  - building on the above, avoiding overreliance on *perceptual* measures regarding robot behavior only – make sure to compare and contrast *objective* measures of robot behavior to identify differences which may not register during user studies.
(3) Being aware of the three different approaches we discuss above that might help to overcome concerns about differences in robot behaviors. We suggest that these considerations are discussed when reporting studies that examine the development of social behaviors on robots.

5 CONCLUSION

The present work describes how we unintentionally developed and deployed gender-biased robot backchanneling behavior. We provide some insights into potential causes - namely, the reductive nature of the use of low-pitch regions as cues for backchanneling, or gender imbalance in the data corpus used for training a deep learning model. Nonetheless, we followed commonly used design protocols for the development of robot social behavior and failed to prevent these biases, which calls for a reflection on these protocols. We discuss potential courses of action that can be taken to address the issue of AI fairness, in a way that is by no means extensive but provides a starting ground for better practices. Especially in the HRI community, where the embodiment of agents adds another layer of social expectations within interactions, these considerations are important points of discussion as we attempt to co-design a more just future for all.

6 STUDY DESIGN AND IMPLEMENTATION DETAILS

Below, we provide a more complete description of the user study and model implementation. A full description of the motivation and results can be seen in Parreira et al. [41].

6.1 Rule-based Model

The **NaïveL** condition is based on the heuristic developed by Ward and Tsukahara [53]. This work suggests a set of rules based on prosodic cues, i.e., pitch. We reproduce the rules below. If all the 5 conditions are met, a BC should be generated:

(P1) *a region of pitch less than the 26th-percentile pitch level and*
(P2) *continuing for at least 110 milliseconds,*
(P3) *coming after at least 700 milliseconds of speech,*
(P4) *providing you have not output backchannel feedback within the preceding 800 milliseconds,*
(P5) *after 700 milliseconds wait.*

In our study, we calculated the pitch distribution of the participant's voice using the YIN estimator [11] over the last 50 s of audio, upon which the percentile level was calculated. Pitch was updated every 10 ms. Backchannel type (non-vocal or vocal) was randomly selected when the five conditions shown above were met. For vocal utterances, the specific sound was selected randomly from a set of pre-defined utterances (*e.g., 'hmm', 'ahh'*). The non-vocal backchannel was realized through a head nod. The "amplitude" - range of the up and down movement - of the nod was randomly sampled from a uniform distribution. During the nod, the robot paused at the lowest point for 0.5 s.

6.2 Data-Driven Model

In this section, we describe the process of implementing and training a deep learning model for backchanneling behavior. Offline performance metrics are provided.

6.2.1 Problem Formulation. The generation of backchanneling in a TAPS session was formulated as a sequential decision-making problem. At any time step t, the environment (user's voice features) is captured as a state variable $s_t \in S$. The robot takes actions $a_t \in A$. The model is divided into two sequential modules, *timing* and *type* of backchannel. It first chooses between a binary output: *performing BC* or *doing nothing*. If *performing BC*, s_t is again used to decide the *type* of BC to perform. The potential actions a_t are *vocal BC* (vocal utterance), *non-vocal BC* (nodding), or *both.*

Dataset: In order to train the model on human-human conversational data, we looked for unscripted, non-topic-bounded interactions that were publicly available. As per these criteria, the dataset used for training was Cardiff's Conversation Database (CCDb)[2]. The CCDb database consists of 30 5-minute dyadic conversations. The 30 interactions are between 16 different speakers (4F, 12M, age range 25-56 years old). We used the subset of conversations that were annotated for facial expressions and utterances to train the listening behavior policy.

A total of 80 minutes of conversational data were used, as we extracted data from the perspective of each participant. We split the individual audio streams into moments where the participant was the *speaker* or the *listener*, and extracted features only in the *listener* portions of the conversation. As training data, we used the audio features from the *speaker* (other participant) as input features, while extracting the annotated backchannels performed by the *listener* participant as ground-truth for how the model should behave. Positive training instances were moments where vocal backchannels or head motion (nodding) were present.

State space: We extracted speech features from the speaker: 13-dimensional mel-frequency cepstrum coefficients (MFCC) and 4-dimensional prosody features, as per prior literature [26, 36, 48]. The

Table 1: Performance metrics (averaged over all validation folds). Both models used augmented data for training.

Module	Model	Hyperparameters	Performance
Timing	*GRU*	*Lookback: 5,activation: sigmoid,* *batch size: 16, dropout: 0.0,* *loss function: focal,optimizer: Adam*	**Macro Accuracy**: 0.95, **Precision**: 0.52, **Recall**: 0.51, **F1**:0.50 **Margin Accuracy**: 0.95, **Precision**: 0.59, **Recall**: 0.76, **F1**:0.65 **BC Prediction Deviation**: 0.83
Type	*GRU*	*Lookback: 10, activation: sigmoid,* *batch size: 32, dropout: 0.2,* *loss function: MSE, optimizer: SGD*	**Macro Accuracy**: 0.64, **Precision**: 0.37, **Recall**: 0.39, **F1**:0.35

MFCC features were 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 34-item feature vector is composed of the mean and standard deviation of each of these features and is normalized. It is generated with a frequency of 2 Hz (one sample every 0.5 s) in the dataset used for training.

Action Space: Both the *DataL* and *NaïveL* robots performed the same backchannel types - vocal utterance, nod of varying amplitude, or a simultaneous combination of both.

6.2.2 *Training the model.*

We explored different model architectures and adjacent techniques such as data augmentation [36].

Data Augmentation: Murray et al. [36] suggest making use of audio data augmentation as a method to improve the robustness of robot listening behaviors. The authors report that users rank the model trained on augmented data higher than a rule-based and a random model. We adopt this method to augment our own dataset, by making use of *masking* techniques in the time and frequency domains. Training instances (audio features) were partially masked in one or both domains, chosen at random. The original and deformed samples were used for training.

Models: We separated our model into two modules, the **Timing** and the **Type** components, according to the two-stage decision-making process. The **Timing** module learns the appropriate timing to perform backchanneling (like the rule-based Ward and Tsukahara [53] model), and the **Type** module decides which type of backchannel - vocal, non-vocal (nodding) or both - the robot should emit. The first module is a binary classification task (*do BC* or *do nothing*). For positive (*'do BC'*) outputs in this module, the second module (a multiclass classification model) decides BC type.

Multiple authors have revealed the potential for the use of neural networks to learn relevant features automatically [34]; more recently, many works leverage models that consider previous internal states, like LSTMs [36, 48], but these architectures are prone to overfitting. Gated Recurrent Units (GRUs) are an alternative to circumvent this problem [21], as they are less complex and usually preferred over LSTMs for small training datasets.

Model Hyperparameter Tuning: We tested different combinations of models and parameters. For both the *Timing* and *Type* modules, we trained single-layer LSTM and GRU models, followed by a dropout layer and a final dense layer. Each model was trained either on the original or the augmented dataset, considering varying previous timesteps, i.e. lookback sizes (5, 10 and 15, respectively 2.5, 5 or 7.5 seconds of data), activation functions (sigmoid, ReLU and softmax), batch sizes (8, 16, 32), dropout rates (0, 0.2 and 0.4), L2 regularization (0.1, 0.01, 0.001, 0.0001), loss functions (focal loss, MSE, binary cross-entropy or hinge loss) and optimizers (SGD and Adam).

Evaluation Metrics: For the two modules, performance was calculated with macro **accuracy** (number of correct predictions over all predictions), **precision** (probability that emitted BCs match the ground-truth data), **recall** (probability that a ground-truth BC is predicted by the model) and **F1-score** (harmonic mean of precision and recall). For the **Timing** module, in line with similar work [12, 48], including that of Ward and Tsukahara [53], we calculated the above metrics considering a tolerance margin of $[-500, 500]$ ms; we also controlled for robot overreaction with a **Backchannel frequency deviation** metric. For each participant i, we considered the relative difference between the predicted number of backchannels (Y^i_{pred}) and the number of backchannels present in the dataset used for testing (Y^i_{true}):

$$\Delta BC^i_{dev} = |Y^i_{true} - Y^i_{pred}|/Y^i_{true} \qquad (1)$$

Models' Selection and Performance: Training of the models was carried out with data split into 8 non-overlapping participant folds, in a 6:1:1 train-validation-test split. Early stopping strategies were deployed for both loss and validation loss, and each model was trained for a maximum of 100 epochs. A dropout layer and L2 regularization are intended to prevent overfitting.

Both models (*Timing* and *Type*) were trained similarly. Hyperparameters were fine-tuned based on macro accuracy and F1-score. The final candidates for each model type (LSTM or GRU, augmented and not augmented training data) were then trained using k-fold cross-validation (on 8 folds). For each module, we deployed the model that was most frequently in the top three best-performing models for each of the metrics mentioned above. Both models are single-layer GRUs trained on the augmented dataset (see Table 1). The **Timing** model outperforms other similar model architectures for BC prediction [36, 48].

ACKNOWLEDGEMENTS

This work was partially funded by grants from 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 Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[2] Andrew J. Aubrey, David Marshall, Paul L. Rosin, Jason Vandeventer, Douglas W. Cunningham, and Christian Wallraven. 2013. Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 277–282. https://doi.org/10.1109/CVPRW.2013.48

[3] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

[4] Frances R. Bilous and Robert M. Krauss. 1988. Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication* 8, 3 (1988), 183–194. https://doi.org/10.1016/0271-5309(88)90016-X Special Issue Communicative Accomodation: Recent Developments.

[5] Peter Blomsma, Gabriel Skantze, and Marc Swerts. 2022. Backchannel Behavior Influences the Perceived Personality of Human and Artificial Communication Partners. *Frontiers in Artificial Intelligence* 5 (2022). https://doi.org/10.3389/frai.2022.835298

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (Dec. 2008), 335–359.

[8] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS). *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* October (2017), 254–262. https://doi.org/10.1145/2909824.3020208

[9] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.

[10] Jenny L Davis, Apryl Williams, and Michael W Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021), 20539517211044808.

[11] Alain de Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111, 4 (2002), 1917–1930. https://doi.org/10.1121/1.1458024 arXiv:https://doi.org/10.1121/1.1458024

[12] I.A. de Kok and Dirk K.J. Heylen. 2012. A Survey on Evaluation Metrics for Backchannel Prediction Models. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. University of Texas, 15–18. null ; Conference date: 07-09-2012.

[13] Iwan de Kok, Dirk Heylen, and Louis-Philippe Morency. 2013. Speaker-Adaptive Multimodal Prediction Model for Listener Responses. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (Sydney, Australia) *(ICMI '13)*. Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/2522848.2522866

[14] Skyla Y. Dudek and James E. Young. 2022. Fluid Sex Robots: Looking to the 2LGBTQIA+ Community to Shape the Future of Sex Robots. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI '22)*. IEEE Press, Sapporo, Hokkaido, Japan, 746–749.

[15] Atsushi Fukayama, Takehiko Ohno, Naoki Mukawa, Minako Sawaki, and Norihiro Hagita. 2002. Messages Embedded in Gaze of Interface Agents — Impression Management with Agent's Gaze. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/503376.503385

[16] Maia Garau, Mel Slater, Simon Bee, and Martina Angela Sasse. 2001. The Impact of Eye Gaze on Communication Using Humanoid Avatars. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 309–316. https://doi.org/10.1145/365024.365121

[17] Victor H. Yngve. 1970. On getting a word in edgewise. *Papers of the Sixth Regional Meeting of Chicago Linguistic Society*, 567–577.

[18] Judith A. Hall and Sarah D. Gunnery. 2013. *21 Gender differences in nonverbal communication*. De Gruyter Mouton, Berlin, Boston, 639–670. https://doi.org/doi:10.1515/9783110238150.639

[19] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2016. Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers. In *INTERSPEECH*. 991–995.

[20] Tom Hitron, Benny Megidish, Etay Todress, Noa Morag, and Hadas Erel. 2022. AI bias in Human-Robot Interaction: An evaluation of the Risk in Gender Biased Robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1598–1605. https://doi.org/10.1109/RO-MAN53752.2022.9900673

[21] Hung-Hsuan Huang, Masato Fukuda, and Toyoaki Nishida. 2019. Toward RNN Based Micro Non-verbal Behavior Generation for Virtual Listener Agents. 53–63. https://doi.org/10.1007/978-3-030-21902-4_5

[22] Lixing Huang and Jonathan Gratch. 2013. Explaining the Variability of Human Nonverbal Behaviors in Face-to-Face Interaction. In *Intelligent Virtual Agents*, Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 275–284.

[23] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual Rapport 2.0. In *Intelligent Virtual Agents*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 68–79.

[24] Andrew Hunt and David Thomas. 2000. *The Pragmatic Programmer: From Journeyman to Master*. Addison-Wesley Longman Publishing Co., Inc., USA.

[25] Nusrah Hussain, Engin Erzin, T. Metin Sezgin, and Yucel Yemez. 2019. Batch Recurrent Q-Learning for Backchannel Generation Towards Engaging Agents. https://doi.org/10.48550/ARXIV.1908.02037

[26] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3411764.3445449

[27] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. 2890–2894. https://doi.org/10.21437/Interspeech.2016-118

[28] Barbara Krahé, Andreas Uhlmann, and Meike Herzberg. 2021. The Voice Gives It Away: Male and female pitch as a cue for gender stereotyping. *Social Psychology* 52, 2 (2021), 101–113. https://doi.org/10.1027/1864-9335/a000441 arXiv:https://doi.org/10.1027/1864-9335/a000441

[29] Kelly Ku and Irene Ho. 2014. Metacognitive strategies that enhance critical thinking. *Metacognition and Learning* 5 (05 2014), 251–267. https://doi.org/10.1007/s11409-010-9060-6

[30] Marianne Latinus and Margot Taylor. 2011. Discriminating Male and Female Voices: Differentiating Pitch and Gender. *Brain topography* 25 (11 2011), 194–204. https://doi.org/10.1007/s10548-011-0207-9

[31] Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal, and Iolanda Leite. 2021. Using Explainability to Help Children UnderstandGender Bias in AI. In *Interaction Design and Children*. 87–99.

[32] Louis-Philippe Morency, Iwan Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20 (01 2010), 70–84. https://doi.org/10.1007/s10458-009-9092-y

[33] Helen Mott and Helen Petrie. 1995. Workplace interactions: Women's linguistic behavior. *Journal of Language and Social Psychology* 14, 3 (1995), 324–336.

[34] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using Neural Networks for Data-Driven Backchannel Prediction: A Survey on Input Features and Training Techniques. In *Human-Computer Interaction: Interaction Technologies*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 329–340.

[35] Anthony Mulac, Karen T Erlandson, W Jeffrey Farrar, Jennifer S Hallett, Jennifer L Molloy, and Margaret E Prescott. 1998. "Uh-huh. What's that all about?" Differing interpretations of conversational backchannels and questions as sources of miscommunication across gender boundaries. *Communication Research* 25, 6 (1998), 641–668.

[36] Michael Murray, Nick Walker, Amal Nanavati, Patricia Alves-Oliveira, Nikita Filippov, Allison Sauppe, Bilge Mutlu, and Maya Cakmak. 2022. Learning Backchanneling Behaviors for a Social Robot via Data Augmentation from Human-Human Conversations. In *Proceedings of the 5th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 164)*, Aleksandra Faust, David Hsu, and Gerhard Neumann (Eds.). PMLR, 513–525. https://proceedings.mlr.press/v164/murray22a.html

[37] Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2013. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces* 7, 1 (2013), 19–28.

[38] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Kenneth Funes Mora, Jean-Marc Odobez, and Joakim Gustafson. 2021. Towards an Engagement-Aware Attentive Artificial Listener for Multi-Party Interactions. *Frontiers in Robotics and AI* 8 (2021), 189. https://doi.org/10.3389/frobt.2021.555913

[39] Y. Okato, K. Kato, M. Kamamoto, and S. Itahashi. 1996. Insertion of interjectory response based on prosodic information. In *Proceedings of IVTTA '96. Workshop on Interactive Voice Technology for Telecommunications Applications*. 85–88. https://doi.org/10.1109/IVTTA.1996.552766

[40] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 1–40.

[41] Maria Teresa Parreira, Sarah Gillet, and Iolanda Leite. 2023. Robot Duck Debugging: Can Attentive Listening Improve Problem Solving? https://doi.org/10.48550/ARXIV.2301.06511

[42] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 110–119.

[43] Ronald Poppe, Khiet P. Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel Strategies for Artificial Listeners. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents* (Philadelphia, PA) *(IVA'10)*. Springer-Verlag, Berlin, Heidelberg, 146–158.

[44] Shahverdi Pourya, Alexander Tyshka, Madeline Trombly, and Wing-Yue Geoffrey Louie. 2022. Learning Turn-Taking Behavior from Human Demonstrations for Social Human-Robot Interactions. https://static1.squarespace.com/static/597e827817bffc7a49fdfe42/t/62c5dd3c6583da2e20eee288/1657134414880/IROS_2022.pdf

[45] Derek Roger and Willfried Nesshoever. 1987. Individual differences in dyadic conversational strategies: A further study. *British Journal of Social Psychology* 26, 3 (1987), 247–255.

[46] Derek B Roger and Andrea Schumacher. 1983. Effects of individual differences on dyadic conversational strategies. *Journal of Personality and Social Psychology* 45, 3 (1983), 700.

[47] Ognjen Rudovic, Meiru Zhang, Bjorn Schuller, and Rosalind Picard. 2019. Multi-Modal Active Learning From Human Data: A Deep Reinforcement Learning Approach. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 6–15. https://doi.org/10.1145/3340555.3353742

[48] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. *Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor: 8th International Workshop on Spoken Dialog Systems*. 247–258. https://doi.org/10.1007/978-3-319-92108-2_25

[49] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. https://doi.org/10.48550/ARXIV.1512.05742

[50] Jane G Stout and Heather M Wright. 2016. Lesbian, gay, bisexual, transgender, and queer students' sense of belonging in computing: An Intersectional approach. *Computing in Science & Engineering* 18, 3 (2016), 24–30.

[51] Yolande Strengers and Jenny Kennedy. 2020. *The Smart Wife: Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot.* MIT Press. Google-Books-ID: 9s7tDwAAQBAJ.

[52] Khiet Phuong Truong, Ronald Walter Poppe, and Dirk K.J. Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. In *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA), 3058–3061. http://www.interspeech2010.jpn.org/ null ; Conference date: 26-09-2010 Through 30-09-2010.

[53] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32, 8 (2000), 1177–1207. https://doi.org/10.1016/S0378-2166(99)00109-5

[54] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education*. Technical Report. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[55] Katie Winkle, Erik Lagerstedt, Ilaria Torre, and Anna Offenwanger. 2022. 15 Years of (Who)Man Robot Interaction: Reviewing the H in Human-Robot Interaction. *J. Hum.-Robot Interact.* (nov 2022). https://doi.org/10.1145/3571718 Just Accepted.

[56] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 29–37. https://doi.org/10.1145/3434074.3446910