Mitigation of gender bias in automatic facial non-verbal behaviors generation

Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, Stephane Ayache

Mitigation of gender bias in automatic facial non-verbal behaviors generation

Alice Delbosc*†
Davi, The Humanizers
Puteaux, France
alice.delbosc@lis-lab.fr

Magalie Ochs
Aix-Marseille Univ, CNRS, LIS
Marseille, France
magalie.ochs@lis-lab.fr

Nicolas Sabouret
Université Paris-Saclay, CNRS, LISN
Orsay, France
nicolas.sabouret@universite-paris-saclay.fr

Brian Ravenet
Université Paris-Saclay, CNRS, LISN
Orsay, France
brian.ravenet@limsi.fr

Stéphane Ayache
Aix-Marseille Univ, CNRS, LIS
Marseille, France
stephane.ayache@lis-lab.fr

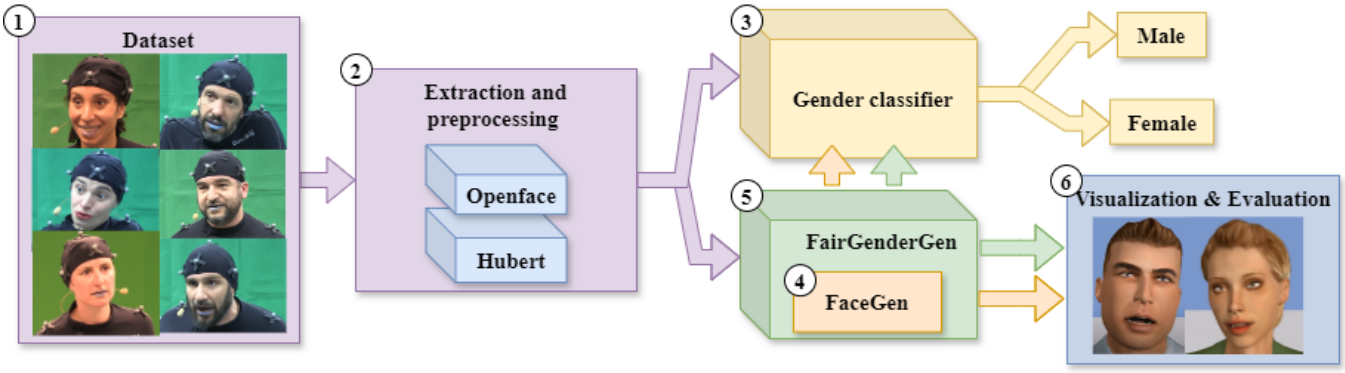

Figure 1: Overview of the fair gender generative model: *(1) From a video corpus, (2) extraction of verbal and non-verbal features. (3) Classifier to verify the feasibility of gender identification based on non-verbal features extracted from the corpus. (4) Evaluation of the gender identification of a facial generation model based on an adversarial approach (FaceGen). (5) Introduction of a model to mitigate the gender bias in facial generation (FairGenderGen). (6) Comparison of the generated behavior of the two models (FaceGen and FairGenderGen) through objective and subjective studies, employing various SIAs.*

ABSTRACT

Research on non-verbal behavior generation for social interactive agents focuses mainly on the believability and synchronization of non-verbal cues with speech. However, existing models, predominantly based on deep learning architectures, often perpetuate biases inherent in the training data. This raises ethical concerns, depending on the intended application of these agents. This paper addresses these issues by first examining the influence of gender on facial non-verbal behaviors. We concentrate on gaze, head movements, and facial expressions. We introduce a classifier capable of discerning the gender of a speaker from their non-verbal cues. This classifier achieves high accuracy on both real behavior data, extracted using state-of-the-art tools, and synthetic data, generated from a model developed in previous work. Building upon this work, we present a new model, *FairGenderGen*, which integrates a gender discriminator and a gradient reversal layer into our previous behavior generation model. This new model generates facial non-verbal behaviors from speech features, mitigating gender sensitivity in the generated behaviors. Our experiments demonstrate that the classifier, developed in the initial phase, is no longer effective in distinguishing the gender of the speaker from the generated non-verbal behaviors.

*Also with Aix-Marseille Univ, CNRS, LIS, Marseille, France.
†Also with Université Paris-Saclay, CNRS, LISN, Orsay, France.

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11…$15.00
https://doi.org/10.1145/3678957.3685732

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Animation**.

KEYWORDS

Non-verbal behavior; behavior generation; bias mitigation; ethics, neural networks; adversarial learning, gradient reversal layer; SIA

ACM Reference Format:
Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, and Stéphane Ayache. 2024. Mitigation of gender bias in automatic facial non-verbal behaviors generation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3678957.3685732

1 INTRODUCTION

Socially Interactive Agents (SIAs) are virtual agents that simulate key properties of face-to-face human conversation, such as verbal and non-verbal behaviors. A number of studies have been carried out to simulate role-playing with SIAs to train one's own skills [4], for example for training doctors to break bad news [31], job interviews [1], negotiation [14], or conflict management [23]. A crucial aspect for the widespread acceptance and use of these applications lies in the believability of the non-verbal behaviors exhibited by the SIAs. The SIAs' non-verbal behavior is particularly important, since several studies underline the positive impact of non-verbal behaviors on knowledge transmission and information retention [7]. In addition, studies indicate that appropriate head movements enhance the overall perception of SIAs, while inappropriate facial expressions can increase their sense of "uncanniness" [37]. Early approaches explored for the automatic generation of SIA's behaviors were based on sets of rules [5, 19]. The rules describe the mapping of words or speech to a facial expression or gesture. These approaches present advantages in terms of communication and control but lack naturalness and variability in behavior generation [30]. Nowadays, most of the research works on behavior generation are based on data-driven approaches [16]. These approaches do not depend on experts in animation and linguistics. They learn the relationships between speech and movements, or facial expressions, directly from data. Among data-driven approaches, deep neural networks have demonstrated their superiority in this task. The commonly employed methodological approach is to extract verbal and non-verbal features from recorded real-world human interactions, and to train a generative model using these real-world datasets [8, 16, 17, 20]. Two key aspects are often evaluated to determine the performance of these models: the human-likeness and the appropriateness of the non-verbal behaviors with speech [21]. However, the presence of possible bias in such models is rarely considered a criterion for evaluating the quality of the model.

Indeed, real-world datasets are often biased [6]. The most frequently identified biases come from key demographic factors like gender. We know, for instance, that men and women differ in their non-verbal behaviors [22]. As generative models learn from our data, most contain biases by simply reproducing those that have been passed to them [11]. This may raise ethical concerns, depending on the intended use of these agents. While it might be wanted to reproduce societal norms and behaviors in SIAs, *e.g.* for better cultural understanding and acceptability, reproducing gender biases can perpetuate harmful stereotypes and inequalities, contributing to the normalization of discriminatory attitudes and behaviors in society. Indeed, a recent study shows that humans inherit the biases of the artificial intelligence they use [38]. Moreover, biased SIAs may make users who don't conform to traditional gender norms or identities feel marginalized or different. Allowing SIAs to perpetuate biases raises ethical questions about the responsibilities of technology creators to promote fairness, equity, and inclusion.

This work addresses this issue of bias and fairness in models of facial non-verbal behavior generation, focusing on gender bias. In the field of generative models, fairness is generally defined as equal generative quality or equal representation [36], for instance of men and women in generated images. In our context, we define fairness as the absence of distinction in the generated non-verbal behaviors, whatever the gender of the speaker. We aim to avoid the perpetuation of gender stereotypes and biases in non-verbal behavior, by adopting an approach in which the generated non-verbal behaviors are not differentiated according to gender. In this article, as a first step, we focus on the gaze, the head movements and the facial expressions. The research questions addressed are:

- Do generative models reproduce potential differences in non-verbal behavior between the genders?
- Can we modify the model to mitigate the gender differences in non-verbal behavior generation without compromising the perceived naturalness and appropriateness of these behaviors with speech?

The paper is organized as follows: we first provide an overview of existing works on fairness in generative models in Section 2. The corpus and feature extraction are presented in Section 3. The baseline generative model is introduced Section 4. The gender classifier and the results of the classification are described in Section 5. Section 6 is devoted to the architecture of the *FairGenderGen* model to generate non-verbal behaviors with mitigation of gender bias. Section 7 is dedicated to the evaluation of the models. Finally, we conclude the paper and introduce perspectives in Section 8. The workflow is illustrated Figure 1.

2 RELATED WORK

Research on bias and fairness has a long history in philosophy, psychology, and in recent years in machine learning [28]. While machine learning ethics often focuses on classification problems, such as gender-neutral hiring [36], recent attention has turned towards the ethical implications of generative models [6, 11, 24].

For these models to be practically viable, they must meet ethical standards and be free from biases that may perpetuate human prejudices. This work contributes to the ethical development of SIAs by addressing gender bias in this domain. To our knowledge, no other research on automatic non-verbal behavior generation has addressed such an ethical dimension.

To effectively rectify these biases and achieve fairness, it is imperative to first establish clear definitions of what constitutes fairness and identify existing discrimination.

2.1 Definition of fairness

The definition of *fairness* varies according to the context in which it is applied. Some definitions of fairness focus on *equal representations* of certain sensible attributes, for example, a generative model that has equal probabilities of producing male or female examples [36].

In our study, we focus on the generation of non-verbal behaviors from speech. While biases in non-verbal behaviors can be addressed by balancing sensitive attributes in datasets, the concept of fairness

related to equal gender representation is not our focus. We wish to concentrate on the intrinsic differences in non-verbal patterns exhibited by individuals of different gender identities.

In the context of generative models, some definitions emphasize *performance fairness* [26]. These approaches seek consistency in generation quality, whatever the sensitive attribute considered, such as gender. Although this approach may apply to our particular situation, generating behaviors with the same performance for men and women in no way ensures that these behaviors do not depend on the gender of the speaker. It is therefore not suitable for working on the generation of non-stereotyped behaviors.

One of the definitions explored in the survey by Mehrabi et al. [28] is: "*an algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process*". We adapt this definition to the generation of non-verbal behaviors and define fairness as "the absence of distinction in the generated non-verbal behaviors, whatever the gender of the speaker". We aim to avoid the perpetuation of gender stereotypes and biases in the non-verbal behavior and to generate non-verbal behaviors that are not differentiated according to gender.

2.2 Approaches to mitigate bias

Biases in model-generated data come mainly from two sources: the dataset and the models themselves. Dataset bias is the main cause of unfairness in generative models. One solution is to work with unbiased data. Practical limitations such as time, resources, and the complex nature of non-verbal behavior, render this approach difficult. We cannot simply balance datasets on the basis of the distribution of sensitive attributes, since bias comes from the fact that individuals have different non-verbal behaviors depending on their gender identity [22].

Models can perpetuate and even amplify biases in the data [6]. Generative Adversarial Networks (GANs), for instance, are trained in an unsupervised way to capture the underlying distribution of the dataset, then generate new data from the same distribution [36].

To address these issues, researchers have explored various techniques, including pre-processing, in-processing, and post-processing [3]. While pre-processing and post-processing methods directly manipulate data, in-processing approaches modify the model during training. Despite advancements in bias mitigation, there is a paucity of research specifically addressing bias in generated non-verbal behaviors. We investigate this aspect from the perspective of generation models in general.

Pre-processing attempts to transform data to remove distribution bias, and post-processing involves modifying the generated data after the model has been trained. Xu et al. [39] work with adversarial networks, trying to generate new data free of the discriminant attribute. They generate new datasets similar to real data that are debiased and preserve good data utility. We felt that it was more effective to operate at the learning stage, building a model that learns from our "biased" data "non-biased" non-verbal behaviors whatever the speaker's gender is.

Several methods have been proposed to mitigate biases in generative models with an in-processing approach. In the context of image generation, Choi et al. [6], Teo et al. [36] use a complementary unbiased dataset as a supervisory signal to detect bias in the baseline data and bring the distribution of the baseline data closer to the reference data. In these works, it is assumed that an unbiased dataset can be accessed or constructed. Zhang et al. [40] employ adversarial learning by presenting a model in which they try to maximize the accuracy of a predictor and at the same time minimize the ability of an adversary to predict the sensitive variable. They use adversarial learning to mitigate sensitive attributes, a method noted by Frankel and Vendrow [11] as costly to train. Frankel and Vendrow [11] develop a method that uses a small neural network ahead of the existing generator to perturb the latent variables. While this approach effectively addresses fairness, it increases both training and inference times due to the additional network layer.

Similarly, some studies seek to learn latent representations that remain invariant with respect to a given variable. One example is the *Variational Fair Autoencoder* [25], which extends the semi-supervised variational autoencoder to acquire representations explicitly invariant to known dataset attributes. By employing a Maximum Mean Discrepancy regularizer, they promote invariant latent variable distributions. This approach necessitates the use of a specialized variational encoder architecture.

The field of domain adaptation, for example the work of Ganin and Lempitsky [12], closely relates to this approach by seeking to minimize the discrepancy between feature distributions of two domains. Their findings demonstrate that adaptation can be integrated into nearly any feed-forward model by adding a small set of standard layers along with a novel gradient reversal layer. Unlike previous methods, this technique enables iterative training, reducing computational costs, and the gradient reversal layer is only active during training, not affecting inference. Furthermore, this approach permanently modifies the latent representation, eliminating the need for an additional neural network before the generator. We propose adapting the approach of Ganin and Lempitsky [12] to mitigate gender bias in non-verbal behavior generation.

3 FACIAL BEHAVIORS CORPUS

Focusing on the automatic generation of facial expressions, head movements and gaze, a corpus that emphasizes facial recordings with a balanced representation of male and female speakers is required. For this purpose, we use the *Trueness* corpus [32].

3.1 Presentation and splitting

Trueness is a corpus of scenes of ordinary discrimination, of sexism and of racism [32]. It also includes interactions between authors of discriminatory behavior and witnesses, attempting to sensitize them by acting out various socio-affective behaviors such as aggression, conciliation or denial. These scenes originate from a French forum theater focused on discrimination, with professional actors trained in this domain. Each scene is divided into two videos, representing the perspectives of the first and second persons in the interaction. An essential quality aspect of the facial non-verbal behaviors is the camera's field of view, carefully maintained to capture only the face and torso.

The dataset is divided into two parts, each recorded separately with different actors. The first part comprises a training set, *SetGen*, and a test set, *TestSet*, used for training and evaluating generative models. The second part, *SetClassif*, is dedicated to train the gender classifier. To ensure dataset diversity and prevent data overlap,

individuals are exclusively included in either *SetGen*, *SetClassif*, or *TestSet*. Specifically, *SetClassif* contains approximately 4 hours and 30 minutes of recordings from two male and two female speakers. *SetGen* includes about 2 hours and 57 minutes of recordings from two male and two female speakers. *TestSet* comprises around 41 minutes of recordings, featuring one male and one female speaker.

3.2 Extraction and processing

We automatically extract behavioral features from the existing videos using *Openface* [2] and speech features using the self-supervised speech model *Hubert* [18].

Behavioral features. *Openface* extracts, among others, 28 features characterizing the head, gaze, and facial behaviors of a person on a video at a frequency of 25 fps (frames per second). The eye gaze position is represented in world coordinates, the eye gaze direction in radians, the head rotation in radians, and 17 facial action units in intensity from 1 to 5 (AU01-02, AU04-07, AU09-10, AU12, AU14-15, AU17, AU20, AU23, AU25-26, AU45) based on the Facial Action Coding System [10]. We point out that these features are designed to capture non-verbal facial behaviors, but do not offer precise lip-synchronization.

To ensure that our model learns from clean, plausible data, we filter out images that have been incorrectly processed by *OpenFace*. These include images in which faces are obscured by hands or hair. We then interpolate the transitions between the remaining images. In addition, two further processing steps are applied to head and gaze features. Firstly, the features are smoothed using a median filter with a window size of 7. Secondly, the head and gaze coordinates are centered to ensure that the SIA is facing the user. Finally, as our focus in this project is solely on generating speaking behaviors (and not listening behaviors), we set the behavioral features to zero when the protagonist is not speaking.

These features, noted $F_b \in \mathbf{R}^{28}$, are used for the training. F_b consists of F_{head}, F_{gaze}, and F_{AU}, representing respectively head movements, gaze orientation, and facial expressions.

Speech features. Drawing on Haque and Yumak [17] work on non-verbal facial behavior generation, we use *Hubert* to extract the speech features. In response to various analyses of different layers of self-supervised speech models [33, 34], we compare the model's objective performances using different layers of *Hubert* (Section 7 for more details on the computation of objective performances), and choose to use the twelfth layer to extract the speech features. In *Hubert*, speech features are extracted at a frequency of 50 fps. The speech features extracted from human speech are noted $F_s \in \mathbf{R}^{1024}$.

Sliding window. Human behaviors are primarily generated by analyzing short segments with a sliding window approach, spanning from seconds to minutes, based on the socio-emotional phenomena studied [29]. We segment the videos into 4-second segments with a 0.4-second overlap. Since the speech data has a frame rate of 50 fps, and the behavior data has a frame rate of 25 fps, we use a speech segment length of 200 frames and a behavior segment length of 100 frames, they are aligned during training.

This segmentation process yields 3590 segments for *SetClassif*, comprising 1429 female, 1459 male, and 702 silent segments (where no speech occurs within the 4-second window). *SetGen* consists of 2940 segments, including 1352 female, 1002 male, and 586 silent segments. *TestSet* comprises 676 segments, divided into 267 female, 344 male, and 65 silent segments. No segment appears in more than one set.

The extracted and processed data, forming *SetClassif*, *SetGen*, and *TestSet*, are part of the *ground truth* data. These datasets underpin all the processes outlined in Figure 1. We develop a behavior generation model using standard techniques and an existing model, *FaceGen* (detailed in Section 4). This model is refined to address gender bias, resulting in *FairGenderGen* (described in Section 6). *SetGen* is employed for training both *FaceGen* and *FairGenderGen*. The gender classifier (described in Section 5) is trained using *SetClassif*. It is used for inference on *ground truth* data from *TestSet*, as well as data generated by both *FaceGen* and *FairGenderGen* (using speech features from *TestSet*).

4 THE *FACE-GEN* MODEL

We aim to generate non-verbal facial behaviors for SIAs while they speak. We can formulate the goal as follows: given a set of speech features $F_s[0:T]$, taken from a particular speech segment at constant frame intervals of length $T = 200$, the goal is to generate the sequence of behaviors $Y_b[0:\frac{T}{2}]$ that a SIA is expected to perform during its speech. The distribution of Y_b must be as close as possible to the one of F_b.

Our model is build upon the work of Delbosc et al. [8], who introduced an open-source framework for the automatic generation of non-verbal facial behaviors using action units. We implemented a number of adjustments, including: the audio features extracted with *Hubert*, the reduction of discriminator capacity, the noise formation, and the model hyperparameters. In this section, we present this transformed architecture[1]. This model will serve as the reference "biased" generative model, which we call *FaceGen*.

4.1 Architecture

Like the original model presented in [8], *FaceGen* adopts the structure of an adversarial encoder-decoder. It is termed "adversarial" because it comprises two modules, a generator and a discriminator, mirroring the architecture of a GAN [13]. The term "encoder-decoder" is employed because the generator operates on the principles of a 1D encoder-decoder. Again, as in the original model, both modules receive speech features $F_s[0:T]$, allowing the discriminator to evaluate the believability of the temporal alignment between behavioral and speech features. Preserving this property of the basic model, the discriminator receives (in addition to *ground truth* and generated examples) examples that help it discriminate between speaking and listening phases. These examples associate features of listening behavior with features of speaking, and vice versa. A simplified architecture is shown in the green frame of Figure 2. To describe each module, we use the following notations: *Conv* and *DoubleConv*. A *Conv* block is composed of a convolution 1D, dropout, batch normalization 1D, and Relu. A *DoubleConv* block is the concatenation of two *Conv* blocks.

The generator. The generator generates data by sampling from a noise z and speech features F_s. The features received by the encoder are not the same as in the original model, so we adapted the architecture, maintaining the main modules. The *encoder* initially

[1] https://github.com/aldelb/FairGenderGen

learns F_s representations using two *Conv* blocks followed by three *DoubleConv* blocks. Each *DoubleConv* is preceded by a maxPool layer. Unlike the original model, this representation is added with noise, not concatenated. The noise is generated by creating two random digits for each channel of F_s representation, and using these values to create a noise matching the length of F_s representation, with transition digits following one another progressively. Following this, three additional *DoubleConv* blocks, each preceded by a maxPool layer, are applied. The output of the encoder constitutes the latent representation of our data. The *decoder* consists of three decoding modules to generate non-verbal behaviors, each associated with an output type with different value intervals: a decoder for head movements, a decoder for eye movements, and a decoder for AUs. They consist of five *DoubleConv* blocks and an upSampling layer before each. It uses skip-connectivity with the corresponding layers of the encoder. It ends with a convolution 1D and a tanh activation layer.

The discriminator. In parallel to the generator, the discriminator learns separate representations for F_s and F_b. These representations are learned with three *Conv* blocks for F_s and two *Conv* blocks for F_b, with a maxPool layer after each block. These representations are then concatenated and processed through one *Conv* block and two linear layers, followed by a sigmoid activation layer. To enhance computational efficiency without compromising performance, we significantly reduced the network architecture compared to the original model based on evaluation results.

4.2 Training

FaceGen is optimized with a Wasserstein loss with gradient penalty [15]. The generator G, with the parameters of the encoder θ_e, and the parameters of the decoders θ_d, is supervised with the following loss function:

$$\mathcal{L}_G(\theta_e, \theta_d) = \mathcal{L}_{gaze}(\theta_e, \theta_d) + \mathcal{L}_{head}(\theta_e, \theta_d) + \mathcal{L}_{AU}(\theta_e, \theta_d)$$

where $\mathcal{L}_{gaze}$, $\mathcal{L}_{head}$ and $\mathcal{L}_{AU}$ are the root mean square errors (RMSEs) of the gaze orientation, head movement, and AUs features.

$$\mathcal{L}_{mod}(\theta_e, \theta_d) = \sum_{t=0}^{\frac{T}{2}-1} (F_{mod}[t] - Y_{mod}[t])^2$$

with $mod \in \{gaze, head, AU\}$. The discriminator D, with the parameters θ_a, is optimized through the adversarial loss function:

$$\begin{aligned} L_{adv}(\theta_e, \theta_d, \theta_a) =& \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(F_s, \tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(F_s, x)] \\ &+ \phi \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(||\nabla_{\hat{x}} D(F_s, \hat{x})||_2 - 1)^2] \end{aligned}$$

with $\mathbb{P}_r$ the *ground truth* distribution and $\mathbb{P}_g$ the generated distribution defined by $\tilde{x} = G(z, F_s)$, $z \sim p(z)$. $\mathbb{P}_{\hat{x}}$, used to calculate the gradient norm, samples uniformly between pairs of points sampled from the data distribution $\mathbb{P}_r$ and the generator distribution $\mathbb{P}_g$, $\hat{x} = (l)F_b + (1-l)Y_b$ with $0 \leq l \leq 1$. We use $\phi = 10$. By integrating adversarial loss with direct supervisory loss, our objective is the following:

$$\mathcal{L}_y(\theta_e, \theta_d, \theta_a) = \mathcal{L}_G(\theta_e, \theta_d) + \beta.\mathcal{L}_{adv}(\theta_e, \theta_d, \theta_a)$$

we set $\beta = 1$. We use Adam optimizer for training, with a learning rate of 10^{-4} for the generator and the discriminator. Our batch is size 32. This model was trained for 1200 epochs on a v100 Nvidia GPU, for approximately 14 hours.

5 INVESTIGATING GENDER BIAS

We assess the presence of gender bias in both *ground truth* and *FaceGen*-generated non-verbal facial behaviors. Following our fairness definition (Section 2.1), a bias is present in non-verbal features if we can identify them as coming from a female or male speaker. For this purpose, we build a *gender classifier*, trained on *SetClassif* (detailed in Section 3). This classifier predicts the speaker's gender based on input non-verbal behavior features, excluding segments of complete silence.

Architecture and training. The gender classifier is a compact neural network composed of two *Conv* blocks (see Section 4), each followed by a maxPool operation. Subsequently, there is a linear layer, a ReLU activation function, another linear layer, and finally a log softmax activation layer. The model is trained using cross-entropy loss and the Adam optimizer with a learning rate of 10^{-3} for 10 epochs.

Evaluation and interpretation. We train the classifier 10 times to capture variability in the training process, such as random weight initialization and optimization algorithm stochasticity. We train it on a large subset of *SetClassif* (1394 female segments and 1423 male segments) and validate its performances on a smaller subset (35 female segments and 36 male segments). The classifier achieved a mean accuracy of 85.92% with a standard deviation of 4.20%.

Randomly selecting one of the trained classifiers, we classified *ground truth* data from *TestSet*. The resulting accuracy is 90.18% (Table 1) with 4 misclassifications out of 267 for female speakers and 56 misclassifications out of 344 for male speakers. These results indicate that non-verbal behaviors extracted from the dataset exhibit discernible gender patterns, suggesting that gender influences the *ground truth* non-verbal behavior. With this established, we can now explore our initial research question: 'Do generative models reproduce potential differences in non-verbal behavior between the genders?'.

We classified data generated by *FaceGen*. The resulting accuracy is 80.69% (Table 1) with 44 misclassifications out of 267 for female speakers and 74 misclassifications out of 344 for male speakers.The influence of the speaker's gender is evident in both the *ground truth* data and those generated by *FaceGen*. This finding answers our first research question, confirming that gender influence persists in automatically generated behaviors, despite being less pronounced than in *ground truth* data. Therefore, we aim to explore our second research question: "Can we modify the [*FaceGen*] model to mitigate the gender differences in non-verbal behavior generation without compromising the perceived naturalness and appropriateness of these behaviors with speech?".

6 THE FAIR-GENDER-GEN MODEL

We introduce a new model called *FairGenderGen*, designed to generate facial non-verbal behaviors from speech, while also aiming to mitigate gender bias by producing behaviors that are independent of the speaker's gender.

6.1 Architecture

The model work with speech features $F_s[0 : T]$ as inputs, and label from the label space $\{female, male, silence\}$. The approach will nevertheless be generic and can handle any labels. We assume the existence of three distributions: $\mathbb{P}_f$, $\mathbb{P}_m$ and $\mathbb{P}_s$, which will be referred as the *Female*, the *Male* and the *Silence* distributions. All distribution are unknown. We don't deal with the *silence* labels as we aim to maintain the *Silence* distribution unchanged.

Our goal is to achieve a latent representation of our data that is invariant with respect to gender, meaning we aim to make the distributions $\mathbb{P}_f$ and $\mathbb{P}_m$ as similar as possible. At training time, we have access to labeled examples from both distributions. Measuring the dissimilarity of the distributions is however non trivial as they are consistently changing during the training process.

Building on prior research presented in Section 2, we propose to adapt the approach of Ganin and Lempitsky [12] to mitigate gender bias through domain adaptation with backpropagation. The proposed architecture includes all the modules of the *FaceGen* model (green in Figure 2); the generator with encoding and decoding parts, and the discriminator, which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by incorporating a gender classifier (orange in Figure 2).

This gender classifier takes as input the latent representation of *FaceGen* data, and classifies them according to the speaker's gender, *male* or *female*. It does not receive the *silent* sequence representations. It is connected to the encoder via a gradient reversal layer. This layer multiplies the gradient by a negative constant during the backpropagation-based training, known as the adaptation factor λ. Similar to the original paper [12], we gradually change the adaptation factor from 0 to 1 during the training process.

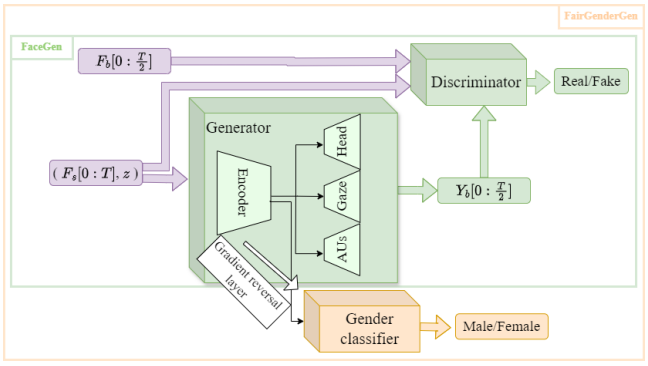

Figure 2: Overall architecture of *FairGenderGen* – *The gender classifier (represented in orange) interacting with FaceGen (represented in green).*

The gender classifier is a small neural network, consisting of two *Conv* blocks (see Section 4), with maxPool after the first block, followed by a linear layer, a ReLU activation function, another linear layer and a log softmax activation layer. Figure 2 illustrates the integration of this classifier with the FaceGen model to form the FairGenderGen model.

Gradient reversal ensures that the distributions over the two genders are made as indistinguishable as possible for the gender classifier, thereby resulting in gender-invariant features. This discriminative classifier is only used during training and does not increase inference time. The modified generator operates identically to the original, except with different outputs, modifying the latent variables for a fair generation.

6.2 Training

To avoid starting from scratch and leverage the learning achieved with the *FaceGen* training, we initialize our discriminator and generator with the *FaceGen* weights.

During the learning stage, we optimize the parameters of the encoder θ_e that maximize the loss of the gender classifier, while simultaneously optimizing the parameters θ_c that minimize the loss of the gender classifier. The gender classifier uses binary cross-entropy as loss function. Otherwise, the training proceeds in a standard manner, minimizing the overall objective L_y with the parameters of the decoders θ_d and the parameters of the discriminator θ_a. By integrating the loss of the gender classifier $\mathcal{L}_{gender}$ and its parameters θ_c, our objective becomes:

$$\mathcal{L}_{fair}(\theta_e, \theta_d, \theta_a, \theta_c) = \mathcal{L}_y(\theta_e, \theta_d, \theta_a) + \alpha.\mathcal{L}_{gender}(\theta_e, \theta_c)$$

with α set to 0.1. We utilize the Adam optimizer for training, with a learning rate of 10^{-4} for the generator and the discriminator. Our batch is size 32. This model was trained for 500 epochs on a v100 Nvidia GPU, requiring approximately 6 hours.

Figure 3 displays the generator's outputs on *TestSet* in three dimensions using *UMAP* visualization [27], with the *Male* and *Female* distributions undeniably closer together for the *FairGenderGen* model. To confirm the visualization results showing that the two distributions are closer together, we conduct an objective and subjective evaluation to assess not only the mitigation of bias, but also the consistency of performances (Section 7).

7 EVALUATION

It is equally important to verify that our non-verbal male and female behaviors are now closer, as it is to ensure that the mitigation of bias has not reduced the quality of the generated behaviors.

To address the first point, we use the gender classifier pretrained on the *SetClassif* data (Section 5). This allows us to assess whether the gender differences in non-verbal behaviors have been minimized in the generated data.

For the second point, we *objectively* and *subjectively* evaluate the model's performances. We compare these metrics with those of the *FaceGen* model to ensure that the quality of the generated non-verbal behaviors has been maintained (maintained, improved or slightly degraded).

7.1 Gender bias

While the gender classifier (Section 5) was able to discriminate between the non-verbal male and female behaviors generated by the *FaceGen* model with an accuracy of 80.69%, its performance significantly dropped to 48.61% when applied to behaviors produced by *FairGenderGen* (Table 1). A closer examination reveals 90 misclassifications out of 267 for female speakers and 224 misclassifications out of 344 for male speakers.

To eliminate gender-based distinctions in generated non-verbal behaviors, the distributions of male and female behaviors were

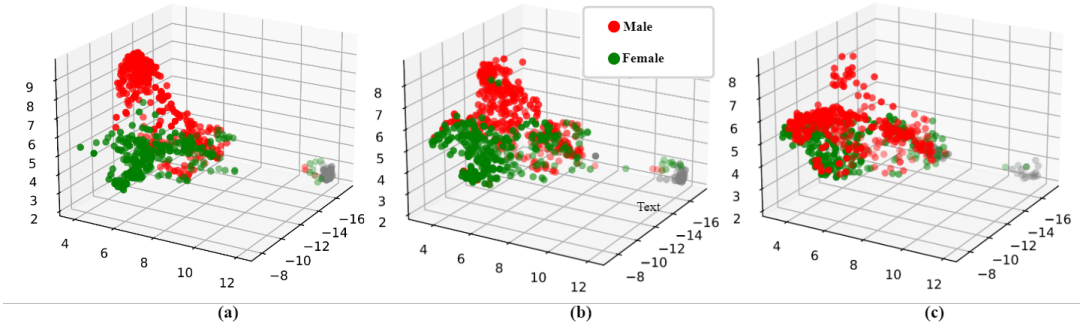

Figure 3: *UMAP* visualization of the non-verbal behaviors - *(a) ground truth behaviors, (b) FaceGen-generated behaviors, and (c) FairGenderGen-generated behaviors. Red and green points represent male and female behaviors, respectively.*

Table 1: Results of the gender classification of *ground truth* behaviors (*Ground truth*), *FaceGen* generated behaviors and *FairGenderGen* generated behaviors – *We report results for all features in terms of accuracy (Acc.) and F1 scores (F1).*

	Ground truth	*FaceGen*	*FairGenderGen*
Acc. / F1	90.18% / 90.21%	80.69% / 80.76%	48.61% / 47.55%

brought closer together. As a consequence, our classifier is now less effective at distinguishing between the two, misclassifying a significant proportion of male behaviors as female. The following section examines the proximity of these distributions and quantifies the performance difference between *FaceGen* and *FairGenderGen*.

7.2 Performance evaluations

To evaluate the *FairGenderGen* model, we generate videos for the two individuals, male and female, who compose the *TestSet*. This involves generating all the segments and averaging overlapping image frames. Our evaluation is based on eight full videos for the objective evaluation (Section 7.3) and four 30-second portions for the subjective evaluation (Section 7.4).

7.3 Objective evaluation

Objective measurements, relying on algorithmic methodologies, provide numerical performance indicators. We use mainly Dynamic Time Warping *DTW*, an algorithm for measuring similarity between two temporal sequences, which may vary in speed.

Distance between males and females. First, *DTW* is employed to assess the similarity between the distributions of male and female non-verbal features across *ground truth* data, *FaceGen*-generated data, and *FairGenderGen*-generated data. For each feature, the *DTW* is computed between the corresponding male and female distributions. An overall gender bias measure is obtained by averaging these *DTW* distances.

Table 2 confirms that the gender bias, *i.e.* the distance between the two distributions, is increased using the *FaceGen* model compared to the *ground truth*. Generative models are capable of amplifying biases existing in the data they are trained on. However, we manage to reduce the distance between these two distributions using the *FairGenderGen* model.

Table 2: *DTW* distance between males and females across *Ground truth*, *FaceGen* and *FairGenderGen* – *The global average.*

	Ground truth	*FaceGen*	*FairGenderGen*
DTW	31.58	32.37	24.70

Distance between the ground truth and the generated behaviors. Second, *DTW* is used to assess the distance between the *ground truth* distributions and the generated distributions. Table 3 indicates that the distributions of *FairGenderGen* are slightly further away from the *ground truth* than those of *FaceGen*, which may lead to a reduction in quality. We add the DTW between a static SIA (central position, AUs at intensity 0) and *ground truth* for additional comparison.

Table 3: *DTW* distance between *ground truth* and a static SIA, generated distributions for *FaceGen* and *FairGenderGen*– *The global average.*

	Static SIA	*FaceGen*	*FairGenderGen*
DTW	29.00	14.18	14.98

While divergence between generated distributions and *ground truth* was expected due to the intended transformation, we have to estimate whether such divergence remains within acceptable limits. Objective measures, while valuable, are insufficient since they neglect the coherence between behaviors and speech, privileging statistical similarity over contextual relevance [21]. Consequently, to evaluate the acceptability of this divergence revealed in objective measures, subjective evaluations play a crucial role.

7.4 Subjective evaluation

To conduct subjective studies, we selected four approximately 30-second speech sequences, two featuring a female and two featuring a male speaker. These sequences were chosen semi-randomly, ensuring coherence in speech over the 30-second duration. Utilizing the *Greta* platform [35], we played these sequences on SIAs, employing a male agent for non-verbal behaviors accompanying a male speech and a female agent for those accompanying a female speech.

For the study setup, we employ the interface of Delbosc et al. [8], inspired by other interfaces widely used in the field of behavior generation. Participants were tasked with evaluating two criteria across the four sequences: believability and temporal coordination with speech of the SIAs' behaviors. Thirty French participants, recruited on social media (15 males, 15 females, mean age 42.7, std 13.4), evaluated the two criteria through direct questions:

- o believability: how human-like do the behaviors appear?
- o temporal coordination: how well does the agent's behavior match the speech? (In terms of rhythm and intonation)

Participants rated each video on a scale from 0 (worst) to 100 (best) for both criteria. The believability criterion was evaluated without sound, while the temporal coordination criterion was evaluated with sound. The results are presented in Table 4. The spectrum of responses reflects variances not just between conditions, but also includes external factors like variations in individual preferences.

Statistical analysis is carried out to examine significant differences between the *FaceGen* and *FairGenderGen* models, but also between male and female behaviors in these models. Initially, the normality of the data is evaluated using the Shapiro-Wilk test, confirming that the data originate from a normally distributed population. Consequently, a repeated ANOVA is used.

Table 4: Results for the believability (Bel.) and coordination (Coo.) criteria – *Average score and standard deviation: mean (std).*

	All	Female	Male
Ground truth Bel.	53.70 (15.39)	**55.27** (15.59)	52.13 (18.76)
FaceGen Bel.	46.31 (15.71)	**49.88** (15.70)	42.73 (19.36)
FairGenderGen Bel.	49.63 (14.65)	45.43 (15.94)	**53.83** (16.46)
Ground truth Coo.	47.42 (16.05)	**51.42** (18.27)	43.42 (15.84)
FaceGen Coo.	43.33 (14.23)	42.85 (16.84)	**43.81** (15.82)
FairGenderGen Coo.	54.22 (17.69)	**54.32** (18.84)	54.13 (18.57)

Perceived believability. There is no evidence of a decline in perceived believability of non-verbal behaviors generated by *FairGenderGen* (Table 4). Statistical analysis indicates no significant difference in perceived believability with *FaceGen* ($p > 0.1$).

However, *FairGenderGen*'s male non-verbal behaviors are significantly rated higher than *FaceGen*'s male non-verbal behaviors ($p = 0.008$). *FairGenderGen* improves the perceived believability of male behavior. Without being significant, *FairGenderGen*'s female non-verbal behaviors tends to be rated lower than *FaceGen*'s female non-verbal behaviors ($p > 0.1$).

In addition, by looking at the contrast in perceived believability between males and females, there is no significant difference in the perceived believability of male and female non-verbal behaviors for the *ground truth* and *FaceGen*. But there is significant differences for *FairGenderGen*'s: where male non-verbal behaviors are rated significantly higher than their female counterparts ($p = 0.029$).

Perceived coordination. *FairGenderGen* is significantly better than *FaceGen* ($p < 0.001$). We also note that, *ground truth* female's behaviors are perceived more coordinated than male's ($p = 0.011$), a difference that disappeared in the generated behaviors for both *FaceGen* and *FairGenderGen* (Table 4). This result shows that there is no decline in performance in terms of coordination of the non-verbal behaviors generated with *FairGenderGen*.

8 DISCUSSION AND FUTURE WORK

Our study highlights a new issue in the field of automatic generation of facial non-verbal behaviors: gender bias. While previous work focused mainly on the believability and coordination of these behaviors with speech, our research highlights the importance of considering the differences in non-verbal behaviors between males and females, differences already observed in real life.

We confirmed, through our analysis with a real-world dataset and the training of a state-of-the-art model in the domain, that gender biases are present in *ground truth* behaviors, as well as in generated behaviors. In this paper, we have proposed a new model, *FairGenderGen*, aiming to mitigate these biases and create non-verbal behaviors independent of the speaker's gender.

Our results show that *FairGenderGen* effectively reduces the gender bias present in the data, even fooling a gender classifier that now recognizes much non-verbal behaviors as female's ones. The subjective evaluation shows that there is no performance loss for this model in terms of perceived coordination. However, our study also reveals a major challenge: the perception of the believability of the generated non-verbal behaviors.

Society has higher expectations of women when it comes to non-verbal behavior. For example, Deutsch et al. [9] revealed that the absence of a smile can be detrimental to a woman's image compared with that of a man, while there is no significant difference in image perception between smiling men and smiling women. Society's expectations of non-verbal behaviors negatively influence the perception of women who don't adopt them.

Our efforts to mitigate gender bias in generated non-verbal behaviors resulted in a notable disparity in perceived believability performance between males and females. Female non-verbal behaviors, generally considered more believable than male ones, became significantly less believable compared to their male counterparts. We believe that these results are due to the higher stereotypical expectations placed on female non-verbal behaviors.

The disparity in perception between males and females raises essential questions for the future of research in this field. Should we direct our efforts towards maintaining stereotypes of non-verbal behaviors to preserve equivalent perceived believability and coordination between men and women, thus reflecting reality? Or should we prioritize an approach aimed at reducing these biases, even at the risk of diminishing the perception of believability?

These reflections are not limited solely to gender biases but could be extended to other sensitive variables such as cultural or racial differences. Exploring these questions more deeply could one day enable us to find answers and develop more equitable and inclusive solutions in the field of automatic generation of non-verbal behaviors.

ACKNOWLEDGMENTS

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2024-AD011014211 on the supercomputer Jean Zay's the V100 partition.

REFERENCES

[1] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment: 10th International Conference, ACE 2013, Boekelo, The Netherlands, November 12-15, 2013. Proceedings 10*. Springer, 476–491.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[3] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: an extensible toolkit for detecting. *Understanding, and Mitigating Unwanted Algorithmic Bias* 2 (2018).

[4] Merijn Bruijnes, Jeroen Linssen, and Dirk Heylen. 2019. Special issue editorial: Virtual Agents for Social Skills Training. , 2 pages.

[5] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[6] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. 2020. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*. PMLR, 1887–1898.

[7] Robert O Davis. 2018. The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review* 24 (2018), 193–209.

[8] Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, and Stéphane Ayache. 2023. Towards the generation of synchronized and believable non-verbal facial behaviors of a talking virtual agent. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 228–237.

[9] Francine M Deutsch, Dorothy LeBaron, and Maury March Fryer. 1987. What is in a smile? *Psychology of Women Quarterly* 11, 3 (1987), 341–352.

[10] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[11] Eric Frankel and Edward Vendrow. 2020. Fair generation through prior modification. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

[12] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[14] Jonathan Gratch, David DeVault, and Gale Lucas. 2016. The benefits of virtual humans for teaching negotiation. In *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16*. Springer, 283–294.

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[16] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.

[17] Kazi Injamamul Haque and Zerrin Yumak. 2023. FaceXHuBERT: Text-less speech-driven E (X) pressive 3D facial animation synthesis using self-supervised speech representation learning. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 282–291.

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[19] Stefan Kopp and Ipke Wachsmuth. 2002. Model-based animation of co-verbal gesture. In *Proceedings of Computer Animation 2002 (CA 2002)*. IEEE, 252–257.

[20] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction* 37, 14 (2021), 1300–1316.

[21] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737* (2023).

[22] Marianne LaFrance and Andrea C Vial. 2016. Gender and nonverbal behavior. (2016).

[23] Minha Lee, Jan Kolkmeier, Dirk Heylen, and Wijnand IJsselsteijn. 2021. Who Makes Your Heart Beat? What Makes You Sweat? Social Conflict in Virtual Reality for Educators. *Frontiers in psychology* (2021), 1365. Publisher: Frontiers.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[25] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

[26] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. 2022. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*. Springer, 344–360.

[27] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[28] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[29] Nora A Murphy and Judith A Hall. 2021. Capturing Behavior in Small Doses: A Review of Comparative Research in Evaluating Thin Slices for Behavioral Measurement. *Frontiers in psychology* 12 (2021), 667326.

[30] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *arXiv preprint arXiv:2301.05339* (2023).

[31] Magalie Ochs, Daniel Mestre, Grégoire De Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Evelyne Lombardo, Daniel Francon, and Philippe Blache. 2019. Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces* 13 (2019), 41–51.

[32] Magalie Ochs, Jean-Marie Pergandi, Alain Ghio, Carine André, Patrick Sainton, Emmanuel Ayad, Auriane Boudin, and Roxane Bertrand. 2023. A forum theater corpus for discrimination awareness. *Frontiers in Computer Science* 5 (2023), 1081586.

[33] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 914–921.

[34] Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[35] Catherine Pelachaud. 2015. Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 5–5.

[36] Christopher TH Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. 2023. Fair generative models via transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2429–2437.

[37] Angela Tinwell, Mark Grimshaw, Debbie Abdel Nabi, and Andrew Williams. 2011. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior* 27, 2 (2011), 741–749.

[38] Lucía Vicente and Helena Matute. 2023. Humans inherit artificial intelligence biases. *Scientific Reports* 13, 1 (2023), 15737.

[39] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*. IEEE, 570–575.

[40] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.