Analyzing gender bias in the non-verbal behaviors of generative systems

Alice Delbosc, Marjorie Armando, Nicolas Sabouret, Brian Ravenet, Stéphane Ayache, Magalie Ochs

Analyzing gender bias in the non-verbal behaviors of generative systems

Alice Delbosc[1,2,3], Marjorie Armando[2,4], Nicolas Sabouret[3], Brian Ravenet[3], Stéphane Ayache[2] and Magalie Ochs[2]

[1]Davi, The Humanizers, Puteaux, France
[2]Aix-Marseille Univ, CNRS, LIS, Marseille, France
[3]Université Paris-Saclay, CNRS, LISN, Orsay, France
[4]Aix-Marseille Univ, CNRS, CRPN, Marseille, France

Abstract

Socially interactive agents (SIAs) simulate essential aspects of human conversation, encompassing both verbal and non-verbal behaviors, and are increasingly integrated into diverse sectors such as healthcare and education. Accurately interpreting and generating non-verbal cues is crucial for enhancing communication effectiveness and user satisfaction. However, the reliance of current research on data-driven approaches in behavior generation for SIAs often results in models inheriting biases from biased real-world datasets, potentially reinforcing societal stereotypes and compromising the ethical integrity of these agents. In this paper, we focus on identifying gender biases in generative models of facial non-verbal behaviors, including gaze, head movements, and facial expressions. By analyzing both real-world interaction data and generated data from a state-of-the-art generative model, and employing a gender classifier, we aim to highlight gender biases present in both types of datasets. The findings from this research initiate discussions on strategies to analyze and mitigate these biases, thereby promoting the development of more inclusive and fair SIAs.

Keywords

Non-verbal behaviors, behavior generation, SIA, virtual agent, ethics, gender bias

1. Introduction

Socially Interactive Agents (SIAs) are virtual agents that simulate key properties of face-to-face human conversation, including both verbal and non-verbal behaviors. With their increasing integration into applications such as healthcare [1] or education [2], it is crucial for these agents to accurately interpret and generate non-verbal behaviors to facilitate effective communication. Non-verbal communication significantly enhances user interaction and satisfaction, making it an essential component of SIA design and implementation [3, 4]. Depending on the methods used to implement non-verbal communication in SIAs, the exhibited behaviors might carry biases, potentially exacerbating issues such as reinforcing societal stereotypes, which undermines the ethical standing of these agents.

Currently, most research on behavior generation in SIAs is based on data-driven approaches [5]. These methods bypass the need for experts in animation and linguistics, instead learning the relationships between speech and movements or facial expressions directly from data. Typically, researchers extract verbal and non-verbal features from recorded real-world human datasets and train generative models using them [6, 7, 8, 9]. However, the presence of possible biases in such models is rarely considered as a criterion for evaluating their quality.

Real-world datasets are often biased [10], frequently due to demographic factors such as gender. Generative models trained on biased data tend to replicate these biases [11]. Moreover, Vicente and Matute [12] show that humans inherit the biases of the artificial intelligence they use. The study revealed that when participants were assisted by a biased AI during a medical diagnostic task, they not only made errors similar to those of the AI, but continued to reproduce these biases even when they subsequently performed the task without assistance. LaFrance and Vial [13] show men and women tend to differ in their non-verbal behaviors. In the context of generative model of non-verbal behaviors, the reproduction of this bias may raise ethical concerns, depending on the intended use of these agents. For example, a generative model that favors high degree of visual dominance in men over women in job interview simulations with SIA may reinforce gender stereotypes. Visual dominance involves *maintaining a relatively higher amount of eye gaze while speaking than while listening and is associated*

REACT'24: viRtual agEnts Against disCriminaTion, Sept 19, 2024, Glasgow, Scotland
alice.delbosc@lis-lab.fr (A. Delbosc); marjorie.armando@univ-amu.fr (M. Armando); nicolas.sabouret@universite-paris-saclay.fr (N. Sabouret); brian.ravenet@limsi.fr (B. Ravenet); stephane.ayache@lis-lab.fr (S. Ayache); magalie.ochs@lis-lab.fr (M. Ochs)

with status and authority. While visual dominance can reduce women's likability and influence, it can actually increase men's influence [14]. This could then have an impact on actual hiring decisions. In this paper, we aim to highlight the presence of such biases and encourage developers of SIA applications to carefully consider their potential effects.

We use a real-world interaction dataset to train a gender classifier and subsequently identify potential biases in the outputs generated by a state-of-the-art model. By basing our analysis in real data, we aim to capture authentic differences in gender interactions, rather than relying solely on synthetic or artificially generated data. After confirming the existence of such biases, we discuss potential research directions and strategies to mitigate them.

The paper is structured as follows: we begin with an overview of biases in non-verbal behavior, followed by a review of current literature on fairness in generative models in Section 2. Section 3 details the gender classifier, dataset, and classification results. In Section 4, we discuss potential avenues for future research stemming from the identification of gender biases in both real non-verbal behaviors and those generated by generative models. Finally, we conclude the paper in Section 5.

2. Related work

Research on gender bias and fairness has a longstanding history in social sciences, and more recently, in machine learning. To effectively address these biases and achieve fairness, it is crucial to identify existing forms of gender discrimination in real-world datasets that could be passed to generative models, and to establish clear definitions of fairness.

2.1. Biases in non-verbal behaviors

2.1.1. In human-human interactions

Biases in non-verbal behaviors can manifest in various forms, often influenced by societal stereotypes and cultural norms. Men and women tend to differ in their non-verbal behaviors [13], with distinct patterns that reflect societal expectations. For instance, societal norms place higher demands on women regarding non-verbal behaviors, such as the detrimental impact on a woman's image for not smiling compared to a man's [15]. Nonverbal dominance is more acceptable for men; maintaining a high degree of visual dominance, associated with status and authority, reduces women's likability and influence but can increase men's influence [14]. Women tend to have more expressive faces, voices, and hands compared to men, who exhibit more restless feet and legs and more open arm and leg postures. During interactions, women tend to gaze, touch, and smile more, standing closer to others except in threatening situations [16].

These biases can be inadvertently encoded into virtual agents through the data used to train the generative models of non-verbal behaviors, or through the designers' own unconscious biases.

2.1.2. In human-SIA interactions

Numerous studies have investigated gender stereotypes through the gendered appearance of virtual agents [17]. The review by Armando et al. [18] shows that gender stereotypes still persist in human-SIA interactions. Female virtual agents are often perceived as less powerful [19], less expert and less knowledgeable [20] than male virtual agents, whereas they are usually seen as more likable [20] and attractive [21].

Even when SIAs show no physical hints about their gender, users still attribute a binary gender according to their non-verbal behaviors, meaning gender stereotypes could still be applied to those agents. McDonnell et al. [22] show that participants perceive the gender of both virtual wooden mannequins and virtual point light agents as ambiguous until they started walking. Their gender is then perceived depending on how participants interpreted their walking motions, just like in the study by Zibrek et al. [23]. Moreover, a female virtual agent is seen as ambiguous when its walking motions are perceived as "male," and the same is true for a male agent with perceived female-walking motions.

By exhibiting non-verbal behaviors that align with gendered societal expectations, virtual agents may receive a positive reception for meeting these expectations. For example, the male agent in the study by Ait Challal and Grynszpan [24] is rated as more agreeable than the female agent in the high direct gaze condition, which goes against the usual results of female agents being perceived as more likable. Wessler et al. [25] replicate backlash effects in human-SIA interactions, a form of social penalty that arises when individuals act counter-stereotypically. In their study, dominant female agents are liked less than all male agents. However, non-verbal behaviors consistent with gender-based societal expectations could also elicit harmful opinions.

Fox and Bailenson [26] study the impact of female agents' sexualized appearance and their gaze behaviors on participants' tendency to express sexist beliefs and rape myth acceptance (misconceptions about rape that place blame on the victim). Among the agents studied, sexist attitudes and acceptance of the rape myth are most expressed towards conservatively dressed agents that avoid eye contact and towards sexualized agents that maintain eye contact. The authors think the representations of female agents might lead people to express more

their sexist attitudes and rape myth acceptance because they trigger existing schemata, align with expectations, and reinforce prevalent stereotypes common in media [27]. Interestingly, unlike previous research with traditional media such as advertisements [28], no main effects of sexualized dress were found on participants' tendency to express sexist attitudes. Instead, the impact of dress needs to be evaluated alongside the agents' non-verbal behaviors. It is important to clarify that examining these interactions should not be interpreted as implying that sexism is a justifiable or legitimate response. Sexist beliefs are not subjective opinions [29] where everyone is entitled to their own view, it is a serious and widespread problem of discrimination. We must be careful not to suggest that sexism is just one viewpoint among many, or that it is somehow acceptable, as this would minimize the real damage it causes.

In contrast to traditional media, interacting with SIAs provides users with a more engaging and immersive experience, similar to real-world social interactions [30, 31]. This interactivity and realism can result in virtual agents having a more significant impact on users' beliefs, attitudes, and behaviors. As highlighted by West et al. [32], the gender bias of interactive systems not only perpetuates stereotypes but also reinforces and extends them. Our goal is to verify the existence of gender differences in non-verbal behaviors found within both the training data and the data generated by the models themselves. Before proceeding, a clear definition of gender equity in non-verbal behaviors is necessary.

2.2. Fairness in generative models

Biases in generative models have been studied across various contexts. Basta et al. [33] investigated gender biases in contextualized word embeddings, Koenecke et al. [34] assessed racial disparities in commercial ASR systems, and Howard et al. [35] explored biases in emotion recognition systems. To our knowledge, there is no existing research studying biases related to virtual agents' non-verbal behaviors by generative models.

The concept of *fairness* in generative models varies depending on the context in which it is applied. Various definitions emphasize *equal representation* of sensitive attributes; for example, ensuring a generative model produces male and female examples equally often [36]. In our study, we focus on generating non-verbal behaviors from speech. Therefore, fairness centered on equal gender representation is not applicable. The generated behaviors are not assigned to a specific gender but may vary based on the speaker's gender, influencing behavioral patterns accordingly. Other definitions emphasize *performance fairness* [37], aiming for consistent generation quality regardless of the sensitive attribute, such as gender. In their survey, Mehrabi et al. [38] discuss a definition stating that "*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, defining fairness as "*the absence of differentiation in the generated non-verbal behaviors regardless of the gender of the speaker*".

Machine learning can be a powerful tool for identifying biases in generative models. Some studies use classification metrics such as accuracy to detect biases [39]. In the realm of non-verbal communication, machine learning algorithms or deep learning models can be trained on datasets of real-world interactions to identify biases. While various classification methods like SVM, neural network classifiers, or random forests can be employed to detect gender biases in non-verbal behaviors, this paper does not aim to conduct a comparative study of classifier performances. Our goal is solely to demonstrate that non-verbal behaviors can be distinguished based on the speaker's gender, without attempting to achieve the highest possible classification performance.

3. Investigating gender bias

This section presents our method to reveal gender biases present in the non-verbal behaviors extracted from a real-world interaction dataset and those generated by a chosen generative model. We work with the extended version of the corpus *Trueness* [40], that focuses on facial recordings with a balanced representation of male and female speakers. We use this real-world data not only to detect biases but, more critically, to train and evaluate a gender classifier. By basing our analysis on authentic interactions, we want to ensure that our classifier identifies gender-related differences in non-verbal behavior, rather than artifacts of synthetic data.

We train a gender classifier using *Trueness*. This classifier analyzes features of non-verbal behaviors to accurately identify the gender of the speaker, distinguishing between male and female speakers. The chosen generative model is *FaceGen*, a model build upon the work of Delbosc et al. [9] and has already been trained with a subset of *Trueness*. *FaceGen* is an open-source model, designed for the automatic generation of non-verbal facial behaviors, including head movements, gaze, and facial expressions.

3.1. *Trueness* dataset

Trueness is a corpus of scenes of ordinary discrimination, of sexism and of racism [40]. It includes interactions between actors simulating discriminatory behaviors and witnesses, attempting to sensitize them by acting out various socio-affective behaviors such as aggression, conciliation or denial. Originating from a French forum theater focused on discrimination, the dataset features trained

professional actors with expertise in this area. Each scene is represented by two separate videos depicting the perspectives of the individuals involved in the interaction. Figure 1 shows images taken during the recording of the dataset. It illustrates the recording methods and the positioning of the actors. The data extracted and processed from this dataset are called the *ground truth* data.

Figure 1: The *Trueness* forum theatre corpus

A part of the dataset is dedicated to train the *FaceGen* model, a second part is used to train the *gender classifier*, and a third part serves as the test set on which we will evaluate gender bias. To ensure dataset diversity and prevent data overlap, individuals are exclusively part of either the first training set, the second training set or the testing set. Specifically, the first training set comprises approximately 4 hours and 30 minutes of recordings, involving 2 male speakers and 2 female speakers. The second training set comprises approximately 2 hours and 57 minutes of recordings, involving 2 male speakers and 2 female speakers. The testing set includes about 41 minutes of recordings, featuring one male speaker and one female speaker.

The behavioral features are automatically extracted from the existing videos using *Openface* [41] and speech features using the self-supervised speech model *Hubert* [42]. *Openface* extracts head position, gaze orientation, and facial expressions, among other features. Eye gaze position is represented in world coordinates, eye gaze direction in radians, head rotation in radians, and facial expressions in intensity from 1 to 5 based on the Facial Action Coding System (FACS) [43] (AU01-02, AU04-07, AU09-10, AU12, AU14-15, AU17, AU20, AU23, AU25-26, AU45). During training and inference, we analyze human behaviors with video segments of 4-second intervals with a 0.4-second overlap.

3.2. Architecture and training

To construct the gender classifier, we employ a neural network architecture composed of two *Conv* blocks. Each *Conv* block processes the input sequence through a structured series of operations. It begins with a 1D convolution, followed by a dropout mechanism, a batch normalization, and finally, a ReLU activation function. Each *Conv* block is followed by a max pooling operation. Following these two *Conv* blocks, there is a linear layer, a ReLU activation function, another linear layer, and finally a log softmax activation layer.

For the training process, we use cross-entropy as loss function. The Adam optimizer is employed to update the model weights, configured with a learning rate of 10^{-3}. The classifier is trained over 10 epochs.

3.3. Evaluation and interpretation

In Table 1, we present the mean accuracy along with the standard deviation. The findings indicate distinct gender patterns in non-verbal behaviors extracted from our dataset. The classifier achieves an average accuracy of 90.11% on the *ground truth* data (testing set). Further analysis reveals 56 segments classified as female speakers out of 344 labeled as male, and 4 segments classified as male speakers out of 267 labeled as female.

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

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

Additionally, the classifier distinguishes between non-verbal behaviors generated by the *FaceGen* model, achieving an accuracy of 80.69%. A closer look at this classification shows 74 segments classified as female speakers out of 344 labeled as male, and 44 segments classified as male speakers out of 267 labeled as female.

Figure 2 displays the test set in three dimensions using *UMAP* visualization [44], highlighting the distributions of *Male* and *Female* data points. The visualization shows that the two distributions are quite different, whether derived from the *ground truth* data or generated by the *FaceGen* model.

The influence of speaker gender is evident in both the *ground truth* data and those generated by *FaceGen*, confirming that gender effects persist in automatically generated behaviors. Recognizing this bias suggests several interesting research directions to explore.

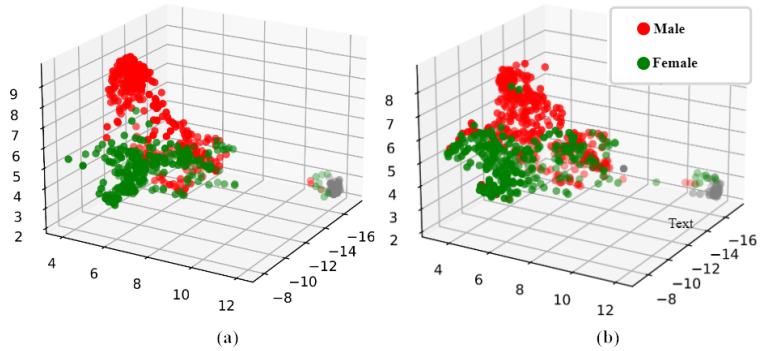

Figure 2: *UMAP* visualization of the non-verbal behaviors - (a) *ground truth* behaviors, (b) *FaceGen*-generated behaviors. Red and green points represent 4-seconds sequences of male and female behaviors, respectively.

4. Discussion on potential research directions

We present research perspectives that we consider pertinent, stemming from the confirmation that automatically generated non-verbal behaviors models replicate gendered non-verbal behavior biases.

Extending the analysis of gender bias

We analyzed the model *FaceGen* for automatic generation of non-verbal behaviors from speech, focusing on facial behaviors such as head movements, gaze direction, and facial expressions via facial action units. It would be beneficial to extend this study to other state-of-the-art models for facial behaviors or explore models generating non-verbal behaviors involving arms, hands, etc.

While our study primarily focused on facial attributes as a whole, each specific behavior—head movements, gaze direction, and facial expressions—merits individual exploration to understand its distinct contribution to gender perception. Analyzing these behaviors separately could provide a more nuanced understanding of how particular non-verbal cues encode or reinforce gender biases. This could be effectively achieved through ablation studies or by employing SHAP [45] to gain insights. Such approaches would help isolate the impact of each non-verbal behavior, shedding light on how they interact and contribute to gender classification.

Another direction to explore is to conduct comparative studies of the gender classifier itself. By exploring various neural network architectures and machine learning models, we could assess their relative effectiveness in detecting non-verbal biases.

Cultural or contextual biases

While we focused on gender biases in non-verbal behaviors, this work can be replicated with adapted datasets to evaluate biases within multicultural datasets. Understanding how non-verbal behaviors manifest across diverse cultural and social contexts offers rich opportunities for exploration. Machine learning techniques can highlight patterns and variations within multicultural datasets, revealing the influence of cultural norms and societal expectations on non-verbal communication.

Approaches for non-verbal bias correction

There are three main ways of debiasing the outputs of models, including pre-processing, in-processing, and post-processing [46]. While pre-processing and post-processing methods directly manipulate data, in-processing approaches modify the model during training. To the best of our knowledge, no research on the automatic generation of non-verbal behaviors has addressed the ethical dimension of the generated behaviors.

Several strategies can be employed to mitigate gender biases in generative models of non-verbal behaviors, drawing from approaches used in other domains. In the context of image generation, Choi et al. [10], Teo et al. [36] use complementary unbiased datasets as supervisory signals during training to reduce bias. Zhang et al. [47] implement adversarial learning to minimize the influence of sensitive attributes. Another approach by Frankel and Vendrow [11] involves introducing a small neural network before the generator to perturb latent variables, effectively addressing bias. Additionally, Louizos et al. [48] extend semi-supervised variational autoencoders to learn representations that are explicitly invariant to certain dataset attributes. These methods provide promising pathways for mitigating biases in the generation of non-verbal behaviors.

Impact on user-agent interactions

Implementing approaches to debias non-verbal behaviors would enable the analysis of how these biases influence

user perception. There are multiple configurations to consider, and various analyses would be relevant. It would be interesting to measure user satisfaction, trust levels, or engagement when interacting with virtual agents that exhibit gendered versus non-gendered behaviors.

Additionally, studying interactions with agents designed to exhibit one gender appearance while displaying non-verbal behaviors associated with another gender can highlight how users interpret and react to these discrepancies. Such a study could reveal insights into how incongruencies influence perceptions of the agent's competence and reliability, and uncover cognitive biases or stereotypes that users apply when faced with conflicting signals.

Long-Term bias evolution

Research by Vicente and Matute [12] demonstrates that humans can inherit biases from the artificial intelligence they interact with. It is crucial to investigate whether biased virtual agents influence stereotypes and potentially alter human non-verbal behaviors over the long term.

Longitudinal studies could also track the evolution of non-verbal biases in virtual agents as the models develop, thereby assessing whether there is an improvement or exacerbation of biases over time.

Moreover, exploring the reproduction of stereotyped behaviors by virtual agents raises questions about their contribution to normalizing discriminatory attitudes and behaviors in society. Biased SIAs might also lead users who do not conform to traditional gender norms or identities to feel marginalized or alienated.

Understanding these dynamics through rigorous longitudinal research is essential for developing strategies to mitigate biases in virtual agents effectively. Such studies could provide insights into how technological advancements can either perpetuate or combat biases, shaping a more equitable future for human-machine interactions.

Need for non-verbal bias reduction

Allowing SIAs to perpetuate biases raises ethical questions about the responsibilities of technology creators to promote fairness, equity, and inclusion. While there may be a desire to replicate societal norms and behaviors in SIAs, such as for better cultural understanding or acceptance, reproducing gender biases could perpetuate harmful stereotypes and inequalities.

This issue underscores the importance of critically evaluating the implications of non-verbal behaviors in SIAs across different domains. Each domain presents unique challenges and opportunities regarding non-verbal biases, making it crucial to consider the specific context in which these agents operate. For instance, non-verbal cues in educational settings might have different impacts compared to those in customer service or healthcare.

Furthermore, the diverse types of users interacting with SIAs add another layer of complexity. Different user groups may perceive and be affected by non-verbal biases in various ways, necessitating a nuanced approach to bias mitigation. This highlights the need for developing contextually appropriate strategies to address non-verbal biases, ensuring that the deployment of SIAs promotes positive and equitable interactions for all users.

Bias in interpreting non-verbal behavior

Considering biases during behavior generation is a crucial initial step towards developing less stereotypical and more inclusive virtual agents. Recent advancements in non-verbal behavior generation models incorporate the interlocutor's behavior into the generation process. This means that how virtual agents respond non-verbally can be influenced by the behavior of the person they are interacting with.

To ensure inclusivity, it is crucial to prevent biases related to the speaker's gender, age, race, and other attributes from influencing how non-verbal behaviors are generated and interpreted. By addressing these biases in the development of virtual agents, we can strive to create interactions that are fair and respectful across diverse user demographics.

5. Conclusion

The identification of non-verbal behavioral biases in SIAs represents an important step toward advancing the understanding and potential reduction of biases in human-machine interactions. We develop a classifier that highlights discernible patterns in non-verbal behaviors, both from the *ground truth* and those generated by the *FaceGen* model, which correlate with gender.

The discussion aims to provide a foundation for future research and development efforts focused on enhancing the inclusivity and equity of virtual agent interactions. By understanding the origins and impacts of these biases, we can explore strategies to mitigate them, thereby promoting the creation of more inclusive and fair SIAs.

While these perspectives are promising, they are not exhaustive, and other directions merit exploration. There are, for example, studies that focus on stereotypes related to the appearance or voice of virtual agents [17, 49, 50]. Świdrak et al. [50] analysed how the degree of perceived masculinity and femininity can influence men's decisions. Integrating these dimensions during the study of non-verbal behaviors would contribute to a holistic approach to bias mitigation in virtual agent design.

References

[1] M. Ochs, D. Mestre, G. De Montcheuil, J.-M. Pergandi, J. Saubesty, E. Lombardo, D. Francon, P. Blache, 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.

[2] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al., 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, 2013, pp. 476–491.

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

[4] A. Tinwell, M. Grimshaw, D. A. Nabi, A. Williams, Facial expression of emotion and perception of the uncanny valley in virtual characters, Computers in Human Behavior 27 (2011) 741–749.

[5] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, M. Neff, A comprehensive review of data-driven co-speech gesture generation, arXiv preprint arXiv:2301.05339 (2023).

[6] T. Kucherenko, D. Hasegawa, N. Kaneko, G. E. Henter, H. Kjellström, Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation, International Journal of Human–Computer Interaction 37 (2021) 1300–1316.

[7] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, C. Theobalt, Learning speech-driven 3d conversational gestures from video, in: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, 2021, pp. 101–108.

[8] K. I. Haque, Z. Yumak, 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, 2023, pp. 282–291.

[9] A. Delbosc, M. Ochs, N. Sabouret, B. Ravenet, S. Ayache, 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, 2023, pp. 228–237.

[10] K. Choi, A. Grover, T. Singh, R. Shu, S. Ermon, Fair generative modeling via weak supervision, in: International Conference on Machine Learning, PMLR, 2020, pp. 1887–1898.

[11] E. Frankel, E. Vendrow, Fair generation through prior modification, in: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), 2020.

[12] L. Vicente, H. Matute, Humans inherit artificial intelligence biases, Scientific Reports 13 (2023) 15737.

[13] M. LaFrance, A. C. Vial, Gender and nonverbal behavior. (2016).

[14] L. L. Carli, Gender and social influence, Journal of Social issues 57 (2001) 725–741.

[15] F. M. Deutsch, D. LeBaron, M. M. Fryer, What is in a smile?, Psychology of Women Quarterly 11 (1987) 341–352.

[16] J. A. Hall, T. G. Horgan, N. A. Murphy, Nonverbal communication, Annual review of psychology 70 (2019) 271–294.

[17] P. Nag, Ö. N. Yalçın, Gender stereotypes in virtual agents, in: Proceedings of the 20th ACM International conference on intelligent virtual agents, 2020, pp. 1–8.

[18] M. Armando, M. Ochs, I. Régner, The impact of pedagogical agents' gender on academic learning: A systematic review, Frontiers in Artificial Intelligence 5 (2022) 862997.

[19] A. L. Baylor, Y. Kim, Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role, in: International Conference on Intelligent Tutoring Systems, Springer, 2004, pp. 592–603.

[20] J. F. Nunamaker, D. C. Derrick, A. C. Elkins, J. K. Burgoon, M. W. Patton, Embodied conversational agent-based kiosk for automated interviewing, Journal of Management Information Sys 28 (2011) 17–48.

[21] R. Lunardo, G. Bressolles, et al., The interacting effect of virtual agents' gender and dressing style on attractiveness and subsequent consumer online behavior, Journal of Retailing and Consumer Services 30 (2016) 59–66.

[22] R. McDonnell, S. Jörg, J. K. Hodgins, F. Newell, C. O'Sullivan, Evaluating the effect of motion and body shape on the perceived sex of virtual characters, ACM Transactions on Applied Perception (TAP) 5 (2009) 1–14.

[23] K. Zibrek, B. Niay, A.-H. Olivier, L. Hoyet, J. Pettré, R. Mcdonnell, The effect of gender and attractiveness of motion on proximity in virtual reality, ACM Transactions on Applied Perception (TAP) 17 (2020) 1–15.

[24] T. Ait Challal, O. Grynszpan, What gaze tells us about personality, in: Proceedings of the 6th International Conference on Human-Agent Interaction, 2018, pp. 129–137.

[25] J. Wessler, T. Schneeberger, L. Christidis, P. Geb-

hard, Virtual backlash: nonverbal expression of dominance leads to less liking of dominant female versus male agents, in: Proceedings of the 22nd ACM international conference on intelligent virtual agents, 2022, pp. 1–8.
[26] J. Fox, J. N. Bailenson, Virtual virgins and vamps: The effects of exposure to female characters' sexualized appearance and gaze in an immersive virtual environment, Sex Roles 61 (2009) 147–157.
[27] C. L. Ridgeway, C. Bourg, Gender as status: an expectation states theory approach. (2004).
[28] M. McKenzie, M. Bugden, A. Webster, M. Barr, Advertising (in) equrality: The impacts of sexist advertising on women's health and wellbeing, Women's Health Issues Paper (2018).
[29] A. C. Curry, G. Abercrombie, Z. Talat, Subjective isms? on the danger of conflating hate and offence in abusive language detection, in: Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), 2024, pp. 275–282.
[30] C. Nass, Y. Moon, Machines and mindlessness: Social responses to computers, Journal of Social Issues 56 (2000) 81–103.
[31] E. Heyselaar, N. Caruana, M. Shin, L. Schilbach, E. S. Cross, Do we really interact with artificial agents as if they are human?, Frontiers in Virtual Reality 4 (2023) 1201385.
[32] M. West, R. Kraut, H. Ei Chew, I'd blush if I could: Closing gender divides in digital skills through education (2019).
[33] C. Basta, M. R. Costa-Jussà, N. Casas, Evaluating the underlying gender bias in contextualized word embeddings, arXiv preprint arXiv:1904.08783 (2019).
[34] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, S. Goel, Racial disparities in automated speech recognition, Proceedings of the National Academy of Sciences 117 (2020) 7684–7689.
[35] A. Howard, C. Zhang, E. Horvitz, Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems, in: 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), IEEE, 2017, pp. 1–7.
[36] C. T. Teo, M. Abdollahzadeh, N.-M. Cheung, Fair generative models via transfer learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 2429–2437.
[37] V. H. Maluleke, N. Thakkar, T. Brooks, E. Weber, T. Darrell, A. A. Efros, A. Kanazawa, D. Guillory, Studying bias in gans through the lens of race, in: European Conference on Computer Vision, Springer, 2022, pp. 344–360.
[38] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.
[39] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira, et al., Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods, Big data and cognitive computing 7 (2023) 15.
[40] M. Ochs, J.-M. Pergandi, A. Ghio, C. André, P. Sainton, E. Ayad, A. Boudin, R. Bertrand, A forum theater corpus for discrimination awareness, Frontiers in Computer Science 5 (2023) 1081586.
[41] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–10.
[42] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, 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.
[43] P. Ekman, W. V. Friesen, Facial action coding system, Environmental Psychology & Nonverbal Behavior (1978).
[44] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[45] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[46] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al., Ai fairness 360: an extensible toolkit for detecting, Understanding, and Mitigating Unwanted Algorithmic Bias 2 (2018).
[47] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.
[48] C. Louizos, K. Swersky, Y. Li, M. Welling, R. Zemel, The variational fair autoencoder, arXiv preprint arXiv:1511.00830 (2015).
[49] C.-P. H. Ernst, N. Herm-Stapelberg, The impact of gender stereotyping on the perceived likability of virtual assistants., in: AMCIS, 2020.
[50] J. Świdrak, G. Pochwatko, A. Insabato, Does an agent's touch always matter? study on virtual midas touch, masculinity, social status, and compliance in polish men, Journal on Multimodal User Interfaces 15 (2021) 163–174.