Getting Emotional Enough: Analyzing Emotional Diversity in Deepfake Avatars

Ilkka Kaate
Marketing, University of Turku
Finland
iokaat@utu.fi

Joni Salminen
University of Vaasa
Finland
jonisalm@uwasa.fi

Soon-Gyo Jung
Qatar Computing Research Institute,
Hamad Bin Khalifa University
Qatar
sjung@hbku.edu.qa

Nina Rizun
Gdansk University of Technology
Poland
nina.rizun@pg.edu.pl

Aleksandra Revina
Department of Economics,
Technische Hochschule Brandenburg
Germany
oleksandra.revina@gmail.com

Bernard J Jansen
Qatar Computing Research Institute,
Hamad Bin Khalifa University
Qatar
jjansen@acm.org

Abstract

When using deepfake technology to represent users, there is a need to convey a reasonable range of emotions to be able to portray different circumstances ranging from positive to negative experiences (e.g., personal struggles). Because it is not known how well deepfake avatars embody emotional diversity, we investigated this aspect among 202 deepfake avatars. Our findings suggest an overall positivity bias in deepfake avatars' emotions. We also found significant gender differences in several emotional expressions, with, male deepfakes scoring higher in "smile" and "calm" emotions, and female deepfake avatars scoring higher in "surprised", "fear", and "happy" emotions. In terms of ethnicity, European and Hispanic deepfake avatars demonstrate the broadest range of "smile", "happy", and "calm" compared to other ethnic groups. Age had no notable bias. No emotion score was normally distributed, suggesting that the range of emotional representativeness among the tested deepfake avatars is skewed. We outline the implications for academics and professionals regarding future development and responsible deployment of deepfake avatars.

CCS Concepts

• **Human-centered computing** → Human computer interaction (HCI).

Keywords

Deepfake avatars, HCI, emotional diversity, user representation

NordiCHI 2024, October 13–16, 2024, Uppsala, Sweden

ACM ISBN 979-8-4007-0966-1/24/10
https://doi.org/10.1145/3679318.3685398

ACM Reference Format:
Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Nina Rizun, Aleksandra Revina, and Bernard J Jansen. 2024. Getting Emotional Enough: Analyzing Emotional Diversity in Deepfake Avatars. In *Nordic Conference on Human-Computer Interaction (NordiCHI 2024), October 13–16, 2024, Uppsala, Sweden.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3679318.3685398

1 Introduction

A deepfake avatar (DA) is a computer-generated representation with a human-like appearance. It is manipulated and managed by a software program to engage, inform, and communicate with users within systems, applications, or services (see examples in Figure 1).

Deepfake technology is often approached in human-computer interaction (HCI) and other studies with a negative undertone, considering it a risk, threat, or harm [10, 20, 66]. However, increasingly, the research community is also recognizing the positive opportunities associated with deepfake technology [31, 33, 34]. Among these is a more efficient representation of end-users using *deepfake avatars* (DAs), which are digital avatars created with the help of deepfake technology. The emergence of deepfake technologies [38, 48, 69] facilitates the incorporation of DAs into information systems that directly engage with real users. To this end, DAs have several potential use cases. For instance, though modern chatbots utilize artificial intelligence (AI) while responding to customers, the chatbots themselves can be quite faceless (or cartoonish) and machine-like [35]. For this purpose, DAs could help to elaborate the user experience (UX) of chatbots and customer services. As a second example, the need for medical personnel in many societies has been increasing due to population aging [12]. There have been advances in AI-driven solutions and robotizing medical and geriatrics staff in, for example, Japan [5, 47]. The reception of such solutions for medical care and for the elderly has been mostly positive, but making such solutions more human-like and humanely responsive could improve their reception [54]. One key element here is the *display of emotion* referring to whether DAs are expressing emotion when interacting with humans–sometimes, for example when DAs interact with people for supportive purposes, designers may wish DAs to display positive emotions; in other occasions, especially when representing user groups that struggle (e.g., with addictions

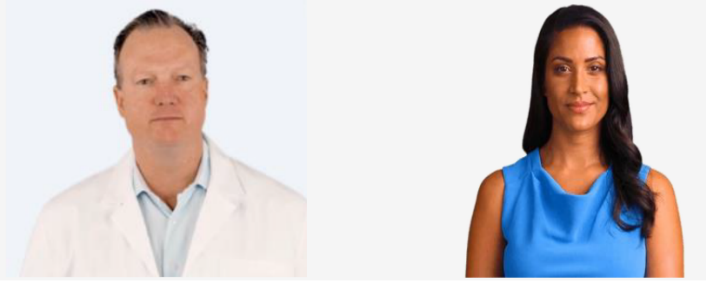

Figure 1: "John" and "Olivia", examples of deepfake avatars from Synthesia.io. These avatars can speak based on a script provided by a human; they can be integrated into a large language model, such as ChatGPT, to respond to user queries in a conversation and thus leveraging a high degree of artificial intelligence. The current study explores the ability of these deepfakes to express *emotions*.

[59], one would expect the DAs to be capable of displaying negative emotions.

Emotions' displaying characteristics of DAs is important for a variety of reasons that are yet to be studied by the scientific community. First, understanding emotions is crucial for enhancing the interaction between humans and DAs. Emotionally expressive DAs improve UX by making interactions more natural and engaging, making DAs more usable [29, 43]. Identifying specific emotions contributing to DAs' diversity helps design more compelling DAs. Second, there could be disadvantageous biases in AI systems. AI often exhibits biases in recognizing and responding to emotions, performing better on certain demographic groups and worse on others [6]. In addition, AI solutions, such as large language models (LLMs), have been found to exhibit human-like biases based on the training data with which the LLM has been trained [1, 72], which can cause hindrances when using deepfake technologies in conjunction with these other AI technologies. Third, affective computing should involve developing systems that recognize and respond to human emotions. Providing data on how DAs mimic human expressions informs the development of more sophisticated algorithms for emotion detection and synthesis [53]. Fourth, ensuring cultural sensitivity in DAs is important for the usability in DAs. Emotions are expressed and interpreted differently across different cultures [9, 13]. Identifying key emotions allows the designing of DAs that effectively communicate with diverse cultural backgrounds, supporting inclusive user-centered design.

As there exists variability in expressing emotions [43], creating DAs that can adapt to this emotional diversity without bias (or skewness) is challenging. This leads to two at least notable knowledge gaps: (1) the effectiveness of DAs in capturing and reflecting the wide range of human emotions, which is crucial for the diverse applications of DAs (ranging across different emotions); and (2) the ability of DAs to represent diverse demographics (and emotional diversity within demographic groups). To address these gaps, our study aims to investigate the diversity and biases in the emotional expressions in DAs. We pose the following research questions (RQs):

RQ1: *What are the key emotional expressions that define the diversity in emotional expressions exhibited by DAs*?

RQ2: *In what ways, if any, do DAs exhibit bias in capturing and reflecting human emotional expressions*?

RQ3: *Are there significant differences in the way DAs express emotions based on age, gender, and ethnic group?*

In this study, our focus is on emotional expressions from static images of DAs. We leave the analysis of facial movement to another study. We selected emotional expressions as the central point of this study due to the recognized significance of emotional expressions in interpersonal communication [18]. Moreover, research has indicated that equipping artificial characters with human-like emotional expressions enhances their perceived human likeness [3, 22]. To address the RQs, we used face detection software *Amazon AWS Rekognition* (AWS) to computationally identify the emotions expressed by the tested 202 DAs that we extracted from three deepfake service providers. AWS was chosen for the study since it has been successfully used in HCI research to detect emotional expressions [30, 42, 56]. We also used two human inspectors (one human inspector who inspected all the data, and the second human inspector validated the results by the first human inspector) to validate the emotional expressions detected by AWS.

Our work has three main practical contributions: (1) *Empirically discovering a remarkable "positivity bias" in DAs* in the emotional expressions exhibited by DAs, with a predominant emphasis on smiles and calm demeanors; (2) A comparison of the way DAs express emotions based on demographic variables, indicating *significant differences in gender, ethnicity, and age, but showing no significant bias in emotional expression*; and (3) *Reflection on the implications of our findings for the future development and responsible deployment of DAs*, prompting inquiry into emotional representation and demographic considerations in the evolution of this technology.

Also, our work brings theoretical insights. Firstly, our work contributes to a better understanding of the range and the nature of emotional expressions that define the diversity in emotional expressions exhibited by DAs. Secondly, we contribute to knowledge about the types of biases that DAs exhibit in capturing and reflecting human emotional expressions. Thirdly, our work contributes to the identification of the demographic factors that influence the diversity in emotional expressions of DAs.

2 Literature Review

2.1 Emotional Expressions in AI Systems

Research from HCI implies that the more a DA looks to be anthropomorphic, the more it seems credible and competent [49]. An anthropomorphic impression of a DA encourages people to become more polite and interactive, and to frequently reciprocate. This enhances the interaction, engagement, and satisfaction of people with DAs [29, 43, 44]. McDuff and Czerwinski [43] emphasize that anthropomorphic DAs, which display a spectrum of human emotions, significantly improve the quality of interaction between the DA and human users, making DAs more relatable and effective. Jack and Schyns [29] emphasize the importance of capturing dynamic emotional expressions, which are fundamental to social communication. In turn, a more human-like appearance of DAs can increase user satisfaction in terms of credibility or goodwill [8, 28]. However, designing and creating a DA that can conduct emotionally apt communication is a challenging task [44]. Studies have explored the universality and cultural variations in emotional expressions, indicating significant differences across different cultural contexts [9, 13, 70]. However, there remains a gap in understanding the full range and nature of emotional emotions that can be exhibited by DAs, especially given the rapid advancements in deepfake technology. It is also a challenge for a DA to communicate with people of different backgrounds and cultures having a limited ability to represent the emotional expressions of diverse people [9, 13]. So, designers and software engineers face the challenge of providing DAs with the ability to represent emotions through emotional expressions to increase user engagement and satisfaction [44].

If the DA has an *emotional intelligence* close to that of a human being, the customers should not see any difference compared to a human operator [37]. Modern office buildings and hotels have reception automation in addition to personnel [40]. Lukanova and Ilieva [40] found that AI-assisted chatbots and robots can enhance the UX of hotel reception encounters, while some restaurants have already outsourced their order reception for robots, chatbots, or at least bare computers [4, 55], and such systems could benefit from deepfakes as well. Pütten et al. [54] show the positive response of people towards those DAs that can collect and analyze human voices and the upper body movements of people to respond accordingly. Expectedly responding to human behavior makes DAs closer to human beings, enhancing system UX among the user base [31, 33, 34]. Emotional diverse DAs need not only account for the positive experiences among the user base but also represent the pain points, challenges, and negative emotions experienced by the user groups they represent [60]. It is common that user representations, such as personas [50], include pain points, concerns, and issues of the user groups the personas represent; not only positive aspirations. AI systems, especially those employing deepfake technology to create DAs, depend on accurately recognizing and reproducing human emotions to ensure effective and engaging interactions [29, 43]. So, understanding cultural differences in emotional expressions is vital for developing AI systems sensitive to the needs of the global audience [9].

To create more inclusive and fair AI systems, deepfake developers ought to use diverse and balanced datasets that represent a wide range of demographic groups and emotional expressions. To this end, Chen and Jack [9] underscore the importance of understanding cultural differences in emotional expression, which is vital for developing AI systems that are sensitive to the needs of a global audience.

HCI and related studies have investigated demographic bias in face detection [21, 67], facial datasets [23], and natural language processing (e.g., word embeddings: [64] and NER [45]), but no study we are aware of investigates demographic bias in DAs. Therefore, this study undertakes a novel, impactful task. Another, perhaps more general, aspect of deepfake research to which this research contributes is the fact that most studies on deepfakes tend to focus on negative effects such as misinformation and other forms of deception (e.g., [14, 39]). However, as pinpointed by Mustak et al. [46], deepfakes also offer opportunities for user-centered design, especially when deployed in an ethically sustainable manner. Therefore, it is integral that there should also be studies focused on the opportunities, and not only the risks, of deepfakes for HCI, which this study addresses.

2.2 Human Emotions Bias in AI Systems

The exploration of emotions in AI systems, particularly DAs, is an important area of research in HCI. Firstly, understanding the range and nature of emotional expressions in DAs is essential for enhancing HCI. Emotionally expressive avatars improve UX by making interactions more natural and engaging [44]. Previous studies have demonstrated that anthropomorphic DAs, which display a wide array of human-like emotions, significantly enhance the quality of interaction between the DAs and humans, making DAs more relatable [29, 43]. Identifying the specific emotions that contribute to the emotional diversity of DAs helps in designing more effective DAs. Secondly, the investigation of biases in emotional generation by DAs is crucial. Commercial AI systems often show gender and racial biases, which extend to emotional recognition [6, 72]. Similarly, Zhao et al. [71] found that individuals rated emotional expressions more positively when associated with high ability in non-threatening contexts, suggesting societal preferences that could influence AI training data.

Biases in AI systems, including DAs, have been a growing concern, particularly regarding how AI systems capture and reflect human emotions. Studies have emphasized the presence of gender and racial biases in AI systems as AI systems tend to perform better on certain demographics than others [6, 32, 57]. Also, these biases can lead to significant discrepancies in the emotional expressions exhibited by DAs, which may affect DAs perceived authenticity and effectiveness. Although previous research has identified bias in AI applications broadly, there is a specific need to investigate how biases affect the emotional expressions of DAs.

Biases in AI in the recognition and generation of human emotional expressions, particularly positivity bias, is a significant concern in affective computing and AI systems [57, 63]. *Positivity bias* refers to the tendency of a group of people in a system to disproportionately favor positive emotions, such as happiness or calmness, over negative emotions, like sadness or anger [43, 71]. Research has indicated that positivity bias can be present in AI systems designed to detect and interpret emotional expressions [6, 43, 71] meaning that AI systems designed to detect and interpret human

emotional expressions detect positive emotional expressions more easily than negative emotional expressions. McDuff and Czerwinski [43] emphasize that emotionally sentient AI systems often exhibit a positivity bias, leading to a skewed detection and representation of human emotions in those systems. Positivity bias can deteriorate the authenticity and effectiveness of AI interactions, as it fails to capture the full spectrum of human emotional expressions[6, 43].

The presence of positivity bias in AI systems is not merely a technical issue but also reflects broader societal preferences for positive emotional expressions, which are often deemed more socially acceptable and desirable [29]. This can materialize in AI systems that are better at recognizing and responding to positive emotions while neglecting or misinterpreting negative emotions. The implications are profound, as positivity bias affects the inclusivity and fairness of AI systems. AI systems exhibiting positivity bias may fail to appropriately respond to users' needs, particularly in contexts where recognizing or expressing negative emotions is crucial, such as mental health support or customer service [57]. Addressing positivity bias in AI systems requires a concerted effort to train AI systems on diverse and balanced datasets that accurately reflect the range of human emotional and emotional expressions. This includes not only ensuring demographic diversity but also representing various emotional states and expressions equally [13].

2.3 Demographic Differences in Emotional Expressions in AI Systems

The influence of demographic factors on the emotional diversity of DAs is a critical area of research, as these factors can significantly impact the effectiveness and relatability of DAs in various applications. Research has demonstrated that cultural and demographic background plays a crucial role in how emotions are expressed and perceived [9, 29, 70] However, there is limited research specifically focusing on how these demographic factors influence the emotional expressions of DAs. Understanding these influences is essential for creating DAs that are not only technologically advanced but also culturally sensitive and inclusive.

Gender plays a crucial role in AI systems' emotional expression detection and generation. Societal stereotypes can lead to women being perceived as more expressive and emotionally positive compared to men [27, 72]. Women are more likely to be seen as happy, while men are more frequently associated with negative emotions such as anger [27]. These gender stereotypes can result in AI systems, trained on gender-biased datasets, which more readily recognize and express positive emotions in women and negative emotions in men, enforcing the positivity bias [6, 72].

Ethnicity also significantly impacts emotional expressions in AI systems. AI systems often perform better on ethnic groups that are mostly represented in AI systems' training datasets, leading to biases where emotional expressions in underrepresented groups are less accurately detected and generated [58, 72]. For example, positive emotions like happiness are more easily recognized in the faces of the ethnic majority within the training data (often white Caucasians [19, 72]), while subtle expressions in ethnic minority groups might be overlooked or misclassified [19, 25, 58, 72]. Similarly, research has also stereotypically found that black ethnicities are generally found more angry in their emotional expressions than white ethnicities [25] which also can affect AI systems' ability to detect and generate emotional expressions across ethnic groups.

Age affects the detection and generation of emotional expressions in AI systems. Older individuals often display different emotional expression dynamics compared to younger individuals [41, 72], which can influence how emotions are perceived and recognized in AI systems related to emotional expressions. Older adults tend to have less intense emotional expressions, which might lead to a lower expression frequency of positive emotions by AI systems trained primarily on data of younger individuals [41]. Consequently, these AI systems might exhibit a positivity bias towards younger individuals, where positive emotions are more easily generated due to more expressive facial features [72].

3 Methods

3.1 Data Collection

To address our RQs, we tabulated data from each DA into summary tables which contained the demographics of age, gender, and ethnic group (a) automatically with a tool called AWS and manually by to human researchers. The initial phase involved the selection of appropriate deepfake providers for the investigation by the lead author. Three service providers were identified for the study: *Synthesia*, *Elai*, and *Colossyan*. These choices were primarily based on the perceived quality of their DAs, which, according to our pilot testing, exhibited a realistic appearance suitable for real-world applications. These providers are marketed to businesses and other client organizations for generating deepfakes for various purposes, including customer service, sales support, online presentations, and even for political endeavors such as pro-government information campaigns [61], though here we focus on user representation. Subsequently, a demographic inspection template for the DAs was established in a spreadsheet. Demographic variables were derived from classifications by the *American Psychological Association (APA)* [36]. The APA demographic classification comprises six ethnic categories: (1) People of African origin, (2) People of Asian origin, (3) People of European origin (white Caucasian), (4) Indigenous Peoples globally (Indigenous peoples of the Americas, Inuit, Australian Aboriginals, Maori), (5) People of Hispanic or Latino ethnicity (later rereferred to as Hispanic in this paper), and (6) People of Middle Eastern origin. Each service provider was individually and manually studied, and each presented DA was inspected manually by the lead author. If a DA was presented, for example, from a different angle or with a different outfit, each variation was considered a unique DA. A screenshot of each DA was saved and meta-tagged after the demographic classification. A total of 202 unique DAs were found on the three providers (February 2023). Screenshots of the DAs and the image nomenclature are provided in the supplementarymaterial. After saving all 202 screenshots, the provider, gender, age, and ethnic information of every DA was entered into a spreadsheet (see Table 1 and Figure 2 for data distributions).

3.2 Emotion Tagging

The DAs' emotional expressions were obtained using AWS, which is a tool for image and video analysis with ML. To automatically analyze DAs' emotional expressions with the AWS, eleven categories were considered: *(a) age, (b) gender, (c) smile, (d) calm, (e) surprised,*

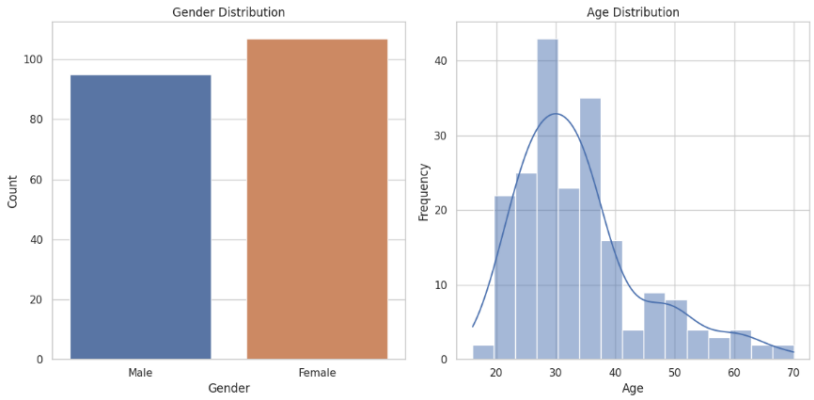

Figure 2: Gender and age distributions in the analyzed deepfake avatars.

Figure 3: Examples of emotion classes given by the AWS. Calm associates with tranquility, peace, and lack of stress [15].

Table 1: Ethnic group distribution.

Ethnic group	N	Relative frequency
European	92	45.6%
Hispanic	34	16.8%
African	37	18.3%
Asian	20	9.9%
Middle Eastern	16	7.9%
Indigenous	3	1.5%

(f) fear, (g) sad, (h) confused, (i) happy, (j) angry, and (k) disgusted. These categories represent emotional expressions in AWS. These emotional expressions were selected for the study since (a) they were available in AWS and (b) the emotional expressions in AWS are based on a subset of emotional expressions categories recognized in a prior study [16]. The prediction of an emotional expression in AWS is "basedonthephysicalappearanceofaperson'sfaceinanimage". AWS gives a confidence value of 0-100 for the estimation of detected emotions c-k (examples in Figure 3). AWS reports values for estimated age as a range from minimum to maximum, from which mean values were calculated for each DA. Gender is reported by AWS as a binary value (male or female), and the estimations of age and gender are based on the physical appearance of a face in an image. AWS reported a confidence value of 0-100 also for the estimation of gender. After the AWS automatic emotional expressions tagging, the ethnic group of each 202 DAs was manually labeled by the lead author using the aforementioned taxonomy. Based on the AWS tagging, the lead author tagged each avatar with one or many of the 11 categories using the same confidence value scale for categories as in AWS giving each avatar a confidence value 0-100 for categories c-k and for age and gender. The confidence value given to the DA by the lead author reflects the amount of an emotion detected by the lead author in the DA.

3.3 Interrater Agreement and Analysis Procedure

We ensured the quality of the human-inspected data labeling via interrater agreement analysis, in which another researcher independently annotated a sample of the data (n=50, i.e., ~25%). The guidance for the annotation process was as follows: "Examine each digital avatar and mark down your findings on (a) gender, (b) age, and (c) ethnic group." The interrater examination revealed that there was high agreement on gender (98.00%, n = 50), age (96.00%, n = 50), and ethnic group (98.00%, n = 50). The disagreed cases were discussed among the researchers, and the final label was assigned based on consensus. Discussion between the researchers followed argumentation from both researchers to the point where a mutual understanding of the originally disagreed case was reached. Two human raters were seen as sufficient for the purposes of the study based on prior studies and guidelines using human raters [24].

Table 2: Mean (M), standard deviation (SD), min, and max values for emotional expressions confidence values observed by AWS.

	Smile	Calm	Surprised	Fear	Sad	Confused	Happy	Angry	Disgusted
Mean	88.01	62.14	7.42	6.53	2.86	11.63	18.87	1.936	1.51
SD	10.06	37.97	3.92	4.97	2.42	23.92	32.33	6.148	5.30
Minimum	52.04	0.08	6.25	5.88	2.15	0.01	0.004	0.01	0.002
Maximum	96.33	99.95	40.11	75.80	27.64	99.17	99.63	62.028	65.50

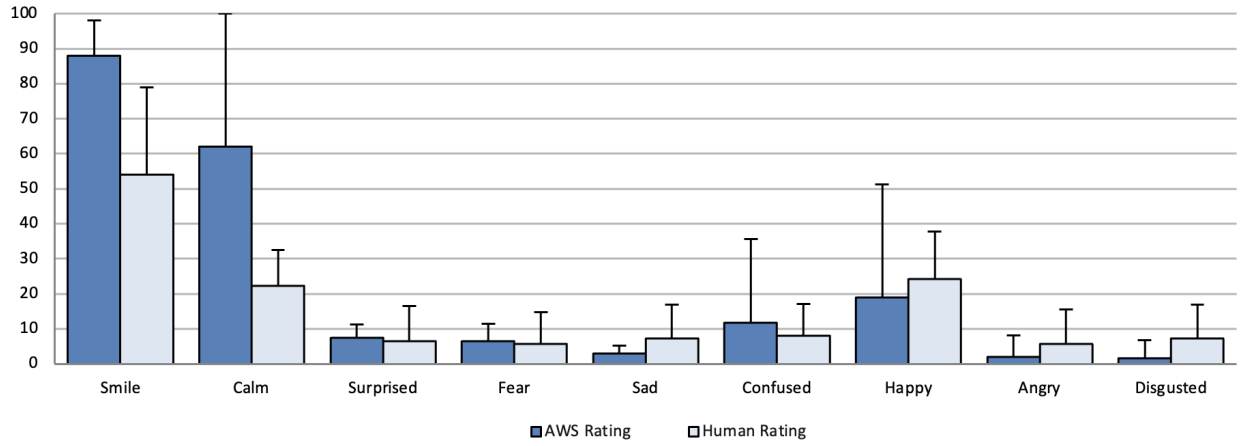

Figure 4: Smile and emotion mean and SDs by AWS (numerical values presented in Table 2) and human rater (numerical values presented in Table 3). Smile and calm are the most common emotions detected by AWS, followed by *happy* and *confused*. *Smile* and *happy* are the most common emotions detected by the human evaluator, followed by *calm* and *confused*. *Surprised* and negative emotions are in the minority. Bars are means, error bars are standard deviations.

For the emotion labeling, another human rater, in addition to the lead author, manually annotated each emotion in the interrater sample (n = 50) using the same confidence value scale as that used by the AWS (0-100) and the lead author. The inter-rater agreement was calculated based on Spearman's rank correlation, indicating that the rank of most observed emotions by the human rater matched well with the most observed emotions by the AWS, $s = 0.9359$, $p < 0.0001$.

To address the RQs, statistical tests were carried out. In these tests, we leverage the variables obtained using AWS (apart from the ethnic group that was manually tagged). Because the data was predominantly non-normally distributed, we applied non-parametric tests, including the Mann-Whitney (assessing whether one of two independent samples (gender in this study) tends to have larger values than the other), Kruskal-Wallis (assessing whether at least one of the groups (ethnic groups in this study) studied tends to have different values from the other groups), and Wilcoxon tests (assessing whether the median differences between pairs of observations (observations in ethnic groups in this study) are different from zero). All three non-parametric tests are suitable to study differences between groups (gender, age, and emotional expressions) in data that is not normally distributed.

4 Findings

4.1 Positivity Bias

To address RQ1, we found that *smile* (M=88.01, SD=10.06) and *calm* (M=62.14, SD=37.97) emotional expressions are the dominant expressions recognized by the AWS, while other emotional expressions are in the minority. The next most common emotional expressions observed by AWS are *happy* and *confused*, but as indicated by the high standard deviations (SD), the range of confidence values output by AWS for these emotional expressions is wide (see Table 2). In other words, there is a high degree of variation among the DAs observed by AWS on the display of emotions with a bias towards happy emotions. Nonetheless, to answer RQ2, the emotions observed from AWS data indicate a "positivity bias" (Figure 4), which refers to the dominance of smiling faces on DAs over other emotional expressions.

The emotional diversity observed by a human evaluator shows a similar smile (M=54.11, SD=24.89) and happy (M=24.21, SD=13.66) are the dominant emotional expressions recognized by the human evaluator. Next, the most common emotional expressions observed by the human evaluator are calm and confused, but as indicated by the high standard deviations (SD), the range of values observed for these emotional expressions is wide (see Table 3). In other words, there is a high degree of variation observed by the human evaluator among the DAs on the display of emotions, with a bias

Table 3: Mean (M), standard deviation (SD), min, and max values for emotions observed by a human evaluator.

	Smile	Calm	Surprised	Fear	Sad	Confused	Happy	Angry	Disgusted
Mean	54.11	22.28	6.46	5.77	7.33	8.02	24.21	5.77	7.15
Standard deviation	24.89	10.16	9.97	9.05	9.50	9.01	13.66	9.70	9.75
Minimum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Maximum	90	60	50	30	30	30	60	40	50

towards happy emotions. The emotional diversity observed by a human evaluator shows a similar positivity bias as seen in the AWS observations (Figure 4). Data in Figure 4 are presented numerically in Table 2 and Table 3.

4.2 The Effect of Age, Gender, and Ethnic Group

This section addresses RQ3. For *age*, we investigated the correlations between age and emotion variables. Only the correlation between age and "confused" reached statistical significance ($p < 0.05$), indicating a weak positive relationship between these variables, $r = 0.1734$, $p = 0.014$. All other correlations were non-significant. So, overall, there appears to be no notable association between the DAs' age and their display of emotions.

For *gender*, we conducted a series of Mann-Whitney U tests. Several significant differences were found:

- Male DAs (M = 89.11) had higher "smile" scores compared to female DAs (M = 87.04), U = 6259.00, p = .0046.
- Male DAs (M = 72.37) also had higher "calm" scores than female DAs (M = 53.06), U = 6755.00, p < .001.
- In turn, female DAs (M = 8.07) had higher "surprised" scores compared to male DAs (M = 6.69), U = 3631.00, p = .0005.
- Also, female DAs (M = 6.90) had higher "fear" scores than male DAs (M = 6.12), U = 4036.00, p = .0117.
- Finally, female DAs (M = 26.27) had higher "happy" scores than male DAs (M = 10.54), U = 3209.00, p < .001.
- The rest of the comparisons were not statistically significant. Figure 5 shows all the gender differences.

For *ethnic groups*, we first carried out Kruskal-Wallis tests for each emotion variable to investigate if the variable scores varied among the ethnic groups. If the Kruskal-Wallis test showed a positive result, we then conducted post-hoc Wilcoxon tests among the ethnic groups. We only report the significant results here. All results, including the non-significant ones, can be found in the supplementary material.

The Kruskal-Wallis test conducted to examine differences in "calm" across different ethnic groups indicated a statistically significant difference, $\chi^2(4) = 21.9166$, $p < .001$. Post-hoc pairwise comparisons using the Wilcoxon test revealed the following results: a significant difference in "calm" between:

- *European and African* DAs, W = -2.3526, p = .0186
- *European and Asian* DAs, W = -2.1462, p = .0319 and
- *European and Middle Eastern* DAs, W = -2.4486, p = .0143.

There was also a significant difference between:

- *African and Hispanic* DAs, W = 3.5248, p < .001
- *Asian and Hispanic* DAs, W = 3.2422, p = .0012 and
- *Hispanic and Middle Eastern* DAs, W = -3.2236, p = .0013.

In general, the *European and Hispanic* DAs exhibited a broader range of "calm" than the other groups, which resulted in their average calmness being lower than for the other groups (see Figure 6). Broader range here refers to the dispersion of detected values for "calm" (Figure 5).

A Kruskal-Wallis test was performed to assess differences in "happy" across the ethnic groups. The test revealed a statistically significant difference, $\chi^2(4) = 17.2542$, $p = .0017$. Further analysis using post-hoc pairwise Wilcoxon tests indicated a significant difference was found between *European and Asian* DAs, W = 2.0396, p = .0414; and *European and Middle Eastern* DAs, W = 2.5375, p = .0112. Also, a significant difference was found between:

- *African and Hispanic* DAs, W = -2.5023, p = .0123
- *Asian and Hispanic* DAs, W = -2.9197, p = .0035 and
- *Hispanic and Middle Eastern* DAs, W = 3.1196, p = .0018.

The *European and Hispanic* (especially the Hispanic) DAs exhibit a broader range of happiness and generally a higher degree of happiness than the other DAs (see Figure 6). Broader range here refers to the dispersion of detected values for "happy" (Figure 5).

Finally, to account for the *simultaneous* effect of *age, gender, and ethnic group* on each of the eleven emotional expressions or emotions outlined at the beginning of Section 3.2, we developed separate regression models for each emotion using the emotion score as a dependent variable and the demographic factors as independent variables. The "calm" regression model explained 19.9% of the variance in "calm" ($R^2 = 0.199$). Among the predictor variables, the ethnic group *European* (β = -19.85, p = 0.009), ethnic group *Hispanic* (β = -34.72, p < 0.001), and gender *Male* (β = 19.38, p < 0.001) showed a significant impact on "calm". The "happy" regression model explained 14.2% of the variance in "happy" ($R^2 = 0.142$). Among the predictor variables, the ethnic group *European* (β = 11.95, p = 0.072), ethnic group *Hispanic* (β = 21.37, p = 0.006), and gender *Male* (β = -13.59, p = 0.003) showed a significant impact on "happy". The other models were ignored due to poor fit. Full test statistics can be found in the online supplementary material.

5 Discussion

5.1 Theoretical Contribution

In terms of RQ1 (*What are the key emotional expressions that define the diversity in emotional expressions exhibited by DAs?*) our research contributes to a better understanding of the wide range and the nature of key emotional expressions that define the diversity in emotional expressions exhibited by DAs [7, 9, 13, 33, 62]. Our research adds to a comprehensive understanding of the range and

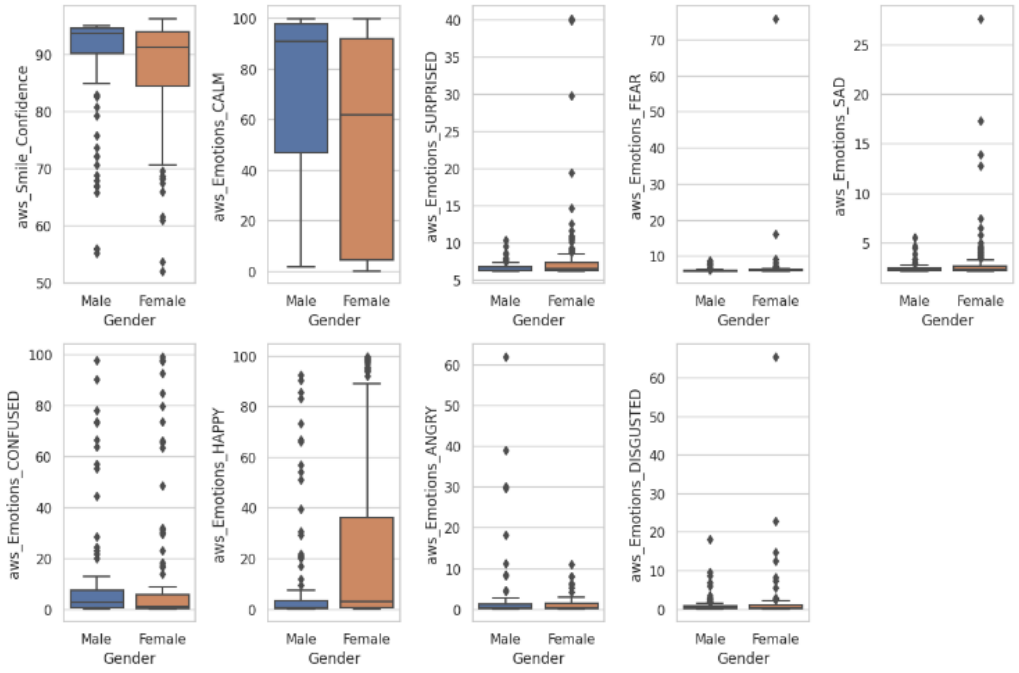

Figure 5: Boxplots showing the emotional differences between deepfake avatars by their gender.

nature of key emotional expressions exhibited by DAs. We identified that emotional expressions such as smile, happy, calm, and confused are the most prominent emotions detected in DAs. Other emotional expressions of surprised, fear, sad, angry, and disgusted (non-positive aspirations commonly present in personas [50]) were observed more often by human observer than AWS, which could be due to such emotions being subtle and dependent on minor cues on the human face. These findings extend the existing literature on emotional expressions in AI, emphasizing the need for AI systems to replicate a broad spectrum of human emotions to enhance emotional diversity [9, 13, 26, 33, 48, 62, 68]. In some occasions it would be beneficial for the DA to be able to exhibit also negative aspirations, e.g., in cases of sad events or when the DA is representing struggle, such as addiction [59]. Previous studies have seen the significance of emotional diversity in HCI, but our study specifically quantifies the prevalence and variability of emotional expressions in DAs, which can conceptually be viewed under the theoretical framework of *emotionally intelligent AI*.

In terms of RQ2 (*In what ways, if any, do DAs exhibit bias in capturing and reflecting human emotional expressions?*), our work contributes to the knowledge and understanding of the types of biases that DAs exhibit in capturing and reflecting human emotional expressions [2, 6, 32, 51, 52, 57]. Our findings reveal a notable positivity bias, with a predominant emphasis on smiles and calm emotional expressions. Both AWS analysis and human observer noted this tendency, indicating a consistent pattern in the emotional representation of DAs. Positivity bias we detected in DAs aligns with societal preferences for positive emotional expressions, yet it challenges the authenticity and diversity of AI-generated emotional expressions [9]. However, the dominance of positive emotions may also raise questions about the authenticity and diversity of emotional expressions conveyed by DAs, as they appear to deviate from the wide spectrum of human emotional experiences, which includes portraying sadness, unhappiness, and even pain. The presence of the positivity bias stresses the importance of developing balanced AI training datasets that represent a full spectrum of human emotions, ensuring AI systems can accurately and fairly recognize and reproduce both positive and negative emotional expressions [6, 71]. By addressing the prevalence of positivity bias, our research adds to the theoretical discourse on ethical AI development and the need for inclusive and representative emotional data in training AI systems.

Regarding RQ3 (*Are there significant differences in the way DAs express emotions based on age, gender, and ethnic group?*), we provide empirical evidence of the demographic influences on DAs' emotional expressions and our research contributes to the broader theoretical framework of *demographic bias* in AI and the development of more equitable AI technologies; an issue raised for example by Hess et al. [27] and Rhue [58]. The major theoretical contribution of our work is to identify the demographic factors that influence the diversity in emotional expressions of DAs adding to the theoretical understanding of how demographic factors influence the emotional expressions of DAs in prior literature [9, 29, 72]. We identified substantial gender differences in emotional expressions of DAs, with male DAs exhibiting higher scores in smile and calm emotions, while female DAs scored higher in surprised, fear, and happy emotions. These gender-specific emotional patterns may reflect (or stem from) societal stereotypes and expectations, which implies that gender representation may reinforce biases when deploying DAs. Additionally, we found ethnic differences, particularly with

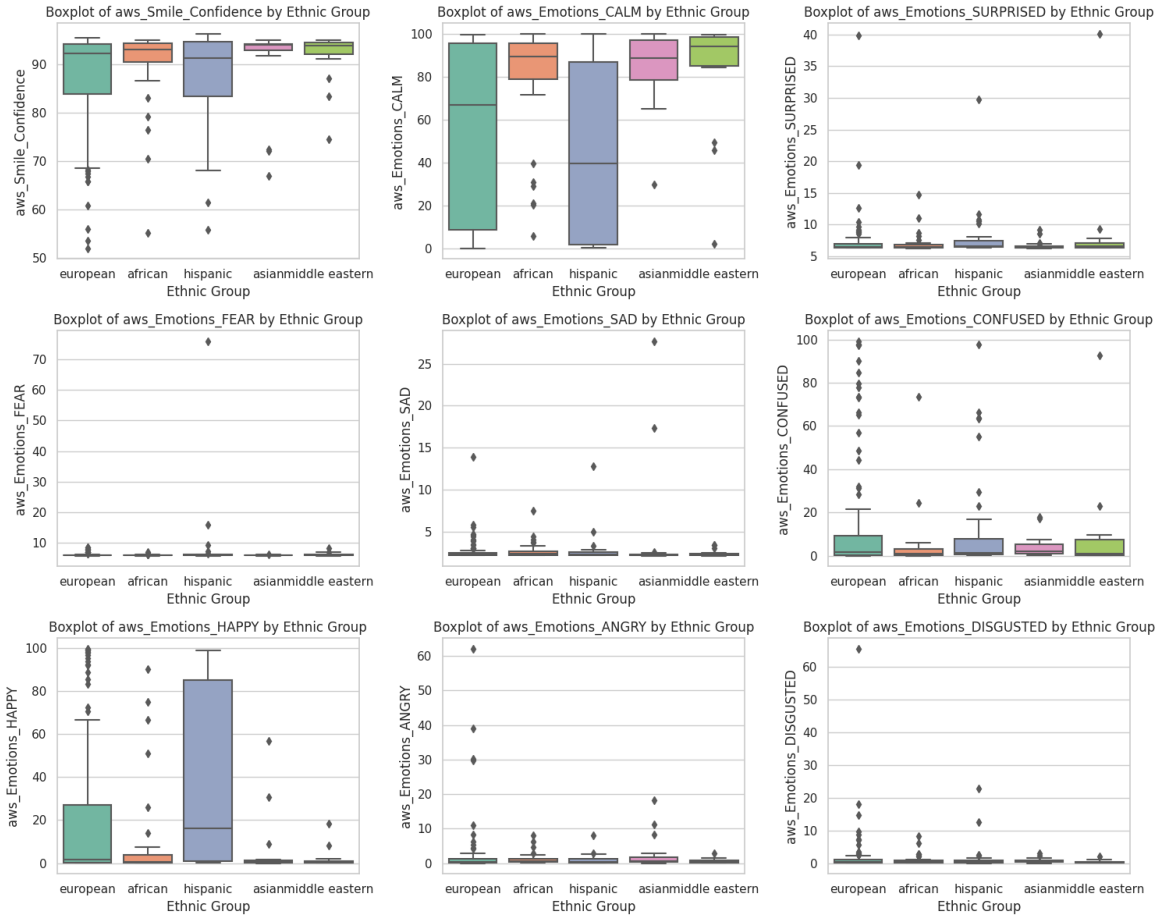

Figure 6: Boxplots of emotion scores from AWS by ethnic group. European and Hispanic ethnic groups seem to portray a range of emotions, which may be a consequence of their larger prevalence in the dataset; 49.0% and 16.8%, respectively.

European and Hispanic DAs displaying a broader range of emotional expressions compared to other ethnic groups. There are also differences in emotional expressions interpretations between different cultures [9, 13, 29] that we might have overlooked due to our reliance on AWS and a Western-based human annotator. The ethnic differences could be explained through the "marketability" of the deepfake technology. That is, the Western world is the major market for these DAs, including television corporations and newspaper offices [65], the automobile industry, the chemical industry, and solar energy [11], and technology giants such as Amazon, Google, and Microsoft [17]. This reasoning emphasizes the need for cultural sensitivity and inclusivity in developing DAs that cater to diverse global audiences.

5.2 Practical Implications

Our results both strengthen and expand existing research on DAs. Our results provide recommendations and guidance aimed at academics, professionals, and DA users. First, developers should focus on employing a wider range of emotional expressions beyond smile and calm demeanors in DAs to address the positivity bias identified in our study. This could enhance the realism and effectiveness of DAs in applications that require a more diverse emotional range, such as customer service or mental health support. By reflecting the emotional complexity of human interactions, DAs can improve user engagement and satisfaction, making interactions feel more natural and engaging. Second, addressing differences in emotional expression based on gender, ethnicity, and age is important in DA applications [70] and our study finds that at the moment the emotional expressions' diversity is not high in DAs. Addressing emotional expressions better involves developing and using inclusive and diverse emotional training datasets to ensure that DAs fairly and accurately represent a broader range of emotional experiences. Third, our findings emphasize the role of emotional representation and demographic factors in the evolution of DAs. This knowledge can inform policymakers and ethicists in developing guidelines that promote responsible and ethical use of DAs, particularly in sensitive areas like mental health support, education, and customer service by offering more holistic and emotionally more expressive DAs to meet the requirements of such sensitive areas.

It may be that the result of lacking emotional expressions variety in some age groups or ethnic groups is a direct consequence of there not being many DAs from said groups. Therefore, the primary means to improve the emotional expressions variety across the range of age groups and ethnic groups is to increase the representation of these groups in the DAs provided.

5.3 Limitations and Future Work

This work has several limitations. First, the study acknowledges a potential Western bias in its analysis, given the reliance on AWS and a Western-based human annotator. Cultural variations in emotional expression interpretation could impact the generalizability of the findings, prompting the incorporation of diverse cultural perspectives in future research. Second, the study focused on a set of eleven emotions supported by AWS, and the possible range of all emotions may not be fully captured by the chosen categories. Future research could explore other emotion dimensions, like arousal in DAs. Third, emotional expressions were primarily analyzed visually, neglecting other modalities such as voice tone or gesture, which are crucial components of human communication. Future studies could consider a multimodal approach to gain a more holistic understanding of emotional representation in DAs. Fourth, AWS has its own biases that need further consideration. Fourth, there were two human raters for the emotional expressions in DAs. More human raters could be used to detect more subtle emotional expressions in DAs. Fifth, there could be spontaneous differences between different cultures and the emotional expressions people from different cultures have. For example, people from other cultures could smile more broadly than people from other cultures, given that the emotional trigger remains the same.

Additionally, further research ideas might address the following directions: First, *investigating how cultural factors influence the interpretation and expression of emotions in DAs.* This could involve collaborative studies with participants from diverse cultural backgrounds to ensure a more inclusive analysis. Second, *investigating how users perceive and trust DAs with varying emotional expressions.* Understanding user preferences and expectations regarding emotional authenticity in DAs can guide improvements in avatar design for enhanced user engagement. Third, *conducting longitudinal studies to track the evolution of emotional representation in DAs over time.* This could provide insights into trends, changes, and improvements in deepfake technology's ability to authentically convey a broad spectrum of emotions. Fourth, *analyzing how the emotional expressions of DAs influence user engagement, satisfaction, and interaction dynamics.* This could provide valuable insights for optimizing emotional design in applications such as customer service, education, and healthcare. Fifth, *focusing on ethical considerations associated with the deployment of DAs*, particularly concerning the reinforcement of stereotypes, biases, and the potential impact on user well-being. This could inform ethical guidelines for developers and organizations utilizing DAs. Sixth, *sampling more service providers to reach a wider sample of DAs from different demographic groups.* Seventh, *using DA videos for facial recognition to gather broader data on different ethnic and gender emotional expressions.* In this study, we focused on static DA images. However, given that the emotion of a DA is expected to be dynamic, adding a dynamic perspective, for example, through DA video analysis, may yield further valuable insights.

6 Conclusion

In this study, our focus was on the emotional expressions of DAs and how the emotional expressions vary between DAs representing different age, gender, and ethnic groups and whether there are any biases in the manner DAs capture human emotional expressions. If DAs do not present a wide range of human emotions, they are likely to be seen as less valuable to be used in design tasks that require empathy or an understanding of difficult situations and human conditions. In our study, we found a happiness bias in DAs, which means that many DAs have a happy face with a smile. This could become problematic on occasions where the avatar is saying things that are not happy, such as talking about death or accidents (or their personal issues). It could result in a mental dissonance for the receiver when the DA has limited emotional sensitivity to the input text sentiment. In the future, it could be possible to create emotionally responsive DAs that can adapt their emotional expressions and tone of voice according to the sentiment of the things it is saying. Overall, the emotional variety of DAs could be improved.

References

[1] Alberto Acerbi and Joseph M. Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proc. Natl. Acad. Sci. U.S.A.* 120, 44 (October 2023), e2313790120. https://doi.org/10.1073/pnas.2313790120

[2] Celia Fernández Aller and Beatriz Peña-Acuña. Facebook and Artificial Intelligence: A Review of Good Practices. In *How Platforms Can Respond to Human Rights Conflicts Online.*

[3] Sahar Aseeri, Sebastian Marin, Richard N. Landers, Victoria Interrante, and Evan S. Rosenberg. 2020. Embodied Realistic Avatar System with Body Motions and Facial Expressions for Communication in Virtual Reality Applications. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, March 2020. IEEE, Atlanta, GA, USA, 580–581. https://doi.org/10.1109/VRW50115.2020.00141

[4] Ashutosh Bhargave, Niranjan Jadhav, Apurva Joshi, Prachi Oke, and SR Lahane. 2013. Digital ordering system for restaurant using Android. *International journal of scientific and research publications* 3, 4 (2013), 1–7.

[5] Olfa Boubaker. 2020. Medical robotics. In *Control Theory in Biomedical Engineering.* Elsevier, 153–204. https://doi.org/10.1016/B978-0-12-821350-6.00007-X

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (*Proceedings of Machine Learning Research*), February 23, 2018. PMLR, 77–91. Retrieved from https://proceedings.mlr.press/v81/buolamwini18a.html

[7] Åsa Cajander, Marta Larusdottir, Elina Eriksson, and Gerolf Nauwerck. 2015. Contextual Personas as a Method for Understanding Digital Work Environments. In *Human Work Interaction Design. Work Analysis and Interaction Design Methods for Pervasive and Smart Workplaces*, José Abdelnour Nocera, Barbara Rita Barricelli, Arminda Lopes, Pedro Campos and Torkil Clemmensen (eds.). Springer International Publishing, Cham, 141–152. https://doi.org/10.1007/978-3-319-27048-7_10

[8] Veena Chattaraman, Wi-Suk Kwon, and Juan E. Gilbert. 2012. Virtual agents in retail web sites: Benefits of simulated social interaction for older users. *Computers in Human Behavior* 28, 6 (November 2012), 2055–2066. https://doi.org/10.1016/j.chb.2012.06.009

[9] Chaona Chen and Rachael E Jack. 2017. Discovering cultural differences (and similarities) in facial expressions of emotion. *Current Opinion in Psychology* 17, (October 2017), 61–66. https://doi.org/10.1016/j.copsyc.2017.06.010

[10] Christopher Clarke, Jingnan Xu, Ye Zhu, Karan Dharamshi, Harry McGill, Stephen Black, and Christof Lutteroth. 2023. FakeForward: Using Deepfake Technology for Feedforward Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581100

[11] Colossyan. 2022. Choose your language. Retrieved December 20, 2022 from https://www.colossyan.com/choose-your-language

[12] Jack M. Colwill, James M. Cultice, and Robin L. Kruse. 2008. Will Generalist Physician Supply Meet Demands Of An Increasing And Aging Population?: Projected shortages could be alleviated if the United States produced four additional generalist graduates in each family and internal medicine residency program each year. *Health Affairs* 27, Suppl1 (January 2008), w232–w241. https://doi.org/10.1377/hlthaff.27.3.w232

[13] Daniel T. Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2018. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* 18, 1 (February 2018), 75–93. https://doi.org/10.1037/emo0000302

[14] Peter R Darke and Robin J.B Ritchie. 2007. The Defensive Consumer: Advertising Deception, Defensive Processing, and Distrust. *Journal of Marketing Research* 44, 1 (February 2007), 114–127. https://doi.org/10.1509/jmkr.44.1.114

[15] Richard J. Davidson. 2004. Well–being and affective style: neural substrates and biobehavioural correlates. *Phil. Trans. R. Soc. Lond. B* 359, 1449 (September 2004), 1395–1411. https://doi.org/10.1098/rstb.2004.1510

[16] Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15 (April 2014). https://doi.org/10.1073/pnas.1322355111

[17] Elai. 2022. List of Supported Languages - Elai.io. Text-to-video AI. Retrieved December 20, 2022 from https://elai.io/elai-languages

[18] Lisa Feldman Barrett. 2021. AI weighs in on debate about universal facial expressions. *Nature* 589, 7841 (January 2021), 202–203. https://doi.org/10.1038/d41586-020-03509-5

[19] Emilio Ferrara. 2023. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* 6, 1 (December 2023), 3. https://doi.org/10.3390/sci6010003

[20] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *CHI Conference on Human Factors in Computing Systems*, 2022. 1–19.

[21] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. 2019. The Harms of Demographic Bias in Deep Face Recognition Research. In *2019 International Conference on Biometrics (ICB)*, June 2019. 1–6. https://doi.org/10.1109/ICB45273.2019.8987334

[22] Nikolaos Gazepidis and Dimitrios Rigas. 2008. Evaluation of Facial Expressions and Body Gestures in Interactive Systems. *International Journal of Computers* 2, 1 (2008), 92–97.

[23] Markos Georgopoulos, James Oldfield, Mihalis A. Nicolaou, Yannis Panagakis, and Maja Pantic. 2021. Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer. *Int J Comput Vis* 129, 7 (July 2021), 2288–2307. https://doi.org/10.1007/s11263-021-01448-w

[24] Natasa Gisev, J. Simon Bell, and Timothy F. Chen. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy* 9, 3 (May 2013), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004

[25] Amy G. Halberstadt, Alison N. Cooke, Pamela W. Garner, Sherick A. Hughes, Dejah Oertwig, and Shevaun D. Neupert. 2022. Racialized emotion recognition accuracy and anger bias of children's faces. *Emotion* 22, 3 (April 2022), 403–417. https://doi.org/10.1037/emo0000756

[26] Jeffrey T. Hancock and Jeremy N. Bailenson. 2021. The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking* 24, 3 (March 2021), 149–152. https://doi.org/10.1089/cyber.2021.29208.jth

[27] Ursula Hess, Sylvie Blairy, and Robert E. Kleck. 2000. The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior* 24, 4 (2000), 265–283. https://doi.org/10.1023/A:1006623213355

[28] Martin Holzwarth, Chris Janiszewski, and Marcus M. Neumann. 2006. The Influence of Avatars on Online Consumer Shopping Behavior. *Journal of Marketing* 70, 4 (October 2006), 19–36. https://doi.org/10.1509/jmkg.70.4.019

[29] Rachael E. Jack and Philippe G. Schyns. 2015. The Human Face as a Dynamic Tool for Social Communication. *Current Biology* 25, 14 (July 2015), R621–R634. https://doi.org/10.1016/j.cub.2015.05.052

[30] Mohanad Azeez Joodi, Muna Hadi Saleh, and Dheya Jassim Khadhim. 2023. Proposed Face Detection Classification Model Based on Amazon Web Services Cloud (AWS). *jcoeng* 29, 4 (April 2023), 176–206. https://doi.org/10.31026/j.eng.2023.04.12

[31] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2023. How Do Users Perceive Deepfake Personas? Investigating the Deepfake User Perception and Its Implications for Human-Computer Interaction. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, September 20, 2023. ACM, Torino Italy, 1–12. https://doi.org/10.1145/3605390.3605397

[32] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, João M. Santos, Essi Häyhänen, Trang Xuan, Jinan Azem, and Bernard J. Jansen. 2024. Modeling the New Modalities of Personas: How Do Users' Attributes Influence Their Perceptions and Use of Interactive Personas? In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, June 27, 2024. ACM, Cagliari Italy, 164–169. https://doi.org/10.1145/3631700.3664882

[33] Ilkka Kaate, Joni Salminen, Joao Santos, Soon-Gyo Jung, Rami Olkkonen, and Bernard Jansen. 2023. The realness of fakes: Primary evidence of the effect of deepfake personas on user perceptions in a design task. *International Journal of Human-Computer Studies* 178, (October 2023), 103096. https://doi.org/10.1016/j.ijhcs.2023.103096

[34] Ilkka Kaate, Joni Salminen, João M. Santos, Soon-Gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2024. "There Is something Rotten in Denmark": Investigating the Deepfake persona perceptions and their Implications for human-centered AI. *Computers in Human Behavior: Artificial Humans* 2, 1 (January 2024), 100031. https://doi.org/10.1016/j.chbah.2023.100031

[35] Mehdi Khosrow-Pour (Ed.). 2015. *Encyclopedia of information science and technology* (Third edition ed.). Information Science Reference, Hershey, PA.

[36] Andy Kiersz. 2015. Most millennials don't identify as millennials. *Most millennials don't identify as millennials*. Retrieved December 13, 2022 from https://www.businessinsider.com/pew-generation-identity-study-2015-9

[37] Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Shreya Bisen, and Vasundhara Rathod. 2018. Implementation of a Chat Bot System using AI and NLP. *IJIRCST* 6, 3 (May 2018), 26–30. https://doi.org/10.21276/ijircst.2018.6.3.2

[38] Yan Li, Manoj a Thomas, and Dapeng Liu. 2021. From semantics to pragmatics: where IS can lead in Natural Language Processing (NLP) research. *European Journal of Information Systems* 30, 5 (September 2021), 569–590. https://doi.org/10.1080/0960085X.2020.1816145

[39] Michael Luca and Georgios Zervas. 2016. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science* 62, 12 (December 2016), 3412–3427. https://doi.org/10.1287/mnsc.2015.2304

[40] Georgina Lukanova and Galina Ilieva. 2019. Robots, Artificial Intelligence, and Service Automation in Hotels. In *Robots, Artificial Intelligence, and Service Automation in Travel, Tourism and Hospitality*, Stanislav Ivanov and Craig Webster (eds.). Emerald Publishing Limited, 157–183. https://doi.org/10.1108/978-1-78756-687-320191009

[41] Carol Z. Malatesta, Michael J. Fiore, and James J. Messina. 1987. Affect, personality, and facial expressive characteristics of older people. *Psychology and Aging* 2, 1 (1987), 64–69. https://doi.org/10.1037/0882-7974.2.1.64

[42] Shraddha Mane and Gauri Shah. 2019. Facial Recognition, Expression Recognition, and Gender Identification. In *Data Management, Analytics and Innovation*, 2019. Springer Singapore, Singapore, 275–290.

[43] Daniel McDuff and Mary Czerwinski. 2018. Designing emotionally sentient agents. *Commun. ACM* 61, 12 (November 2018), 74–83. https://doi.org/10.1145/3186591

[44] Fred Miao, Irina V. Kozlenkova, Haizhong Wang, Tao Xie, and Robert W. Palmatier. 2022. An Emerging Theory of Avatar Marketing. *Journal of Marketing* 86, 1 (January 2022), 67–90. https://doi.org/10.1177/0022242921996646

[45] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. https://doi.org/10.48550/arXiv.2008.03415

[46] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. 2023. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research* 154, (January 2023), 113368. https://doi.org/10.1016/j.jbusres.2022.113368

[47] Yuriko Nakaoku, Soshiro Ogata, Shunsuke Murata, Makoto Nishimori, Masafumi Ihara, Koji Iihara, Misa Takegami, and Kunihiro Nishimura. 2021. AI-Assisted In-House Power Monitoring for the Detection of Cognitive Impairment in Older Adults. *Sensors* 21, 18 (September 2021), 6249. https://doi.org/10.3390/s21186249

[48] Pramukh Nanjundaswamy Vasist and Satish Krishnan. 2022. Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *CAIS* 51, (2022), 590–636. https://doi.org/10.17705/1CAIS.05126

[49] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *J Social Isssues* 56, 1 (January 2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[50] Lene Nielsen, Kira Storgaard Hansen, Jan Stage, and Jane Billestrup. 2015. A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *International Journal of Sociotechnology and Knowledge Development* 7, 1 (2015), 45–61. https://doi.org/10.4018/ijskd.2015010104

[51] Anika Nissen, Colin Conrad, and Aaron Newman. 2023. Are You Human? Investigating the Perceptions and Evaluations of Virtual Versus Human Instagram Influencers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023. ACM, Hamburg Germany, 1–14. https://doi.org/10.1145/3544548.3580943

[52] David Perrett. 2022. Representations of facial expressions since Darwin. *Evolut. Hum. Sci.* 4, (2022), e22. https://doi.org/10.1017/ehs.2022.10

[53] Rosalind W. Picard. 1997. *Affective computing*. The MIT Press, Cambridge, MA, US.

[54] Astrid M. von der Pütten, Nicole C. Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior* 26, 6 (November 2010), 1641–1650. https://doi.org/10.1016/j.chb.2010.06.012

[55] Muhammad Awais Qasim, Faisal Abrar, Sarosh Ahmad, and Muhammad Usman. 2022. AI-Based Smart Robot for Restaurant Serving Applications. In *AI and IoT*

for Sustainable Development in Emerging Countries, Zakaria Boulouard, Mariya Ouaissa, Mariyam Ouaissa and Sarah El Himer (eds.). Springer International Publishing, Cham, 107–123. https://doi.org/10.1007/978-3-030-90618-4_5

[56] Gregorius Rafael, Hendra Kusuma, and Tasripan. 2020. The Utilization of Cloud Computing for Facial Expression Recognition using Amazon Web Services. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, November 17, 2020. IEEE, Surabaya, Indonesia, 366–370. https://doi.org/10.1109/CENIM51130.2020.9297974

[57] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, January 27, 2019. ACM, Honolulu HI USA, 429–435. https://doi.org/10.1145/3306618.3314244

[58] Lauren Rhue. 2018. Racial Influence on Automated Perceptions of Emotions. *SSRN Journal* (2018). https://doi.org/10.2139/ssrn.3281765

[59] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, May 11, 2024. ACM, Honolulu HI USA, 1–20. https://doi.org/10.1145/3613904.3642036

[60] Joni Salminen, Sercan Şengün, João M. Santos, Soon-Gyo Jung, and Bernard Jansen. 2022. Can Unhappy Pictures Enhance the Effect of Personas? A User Experiment. *ACM Trans. Comput.-Hum. Interact.* 29, 2 (January 2022), 14:1-14:59. https://doi.org/10.1145/3485872

[61] Adam Satariano and Paul Mozur. 2023. The People Onscreen Are Fake. The Disinformation Is Real. *The New York Times*. Retrieved February 8, 2023 from https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

[62] Girish Kumar Solanki and Anastasios Roussos. 2023. Deep Semantic Manipulation of Facial Videos. In *Computer Vision – ECCV 2022 Workshops*, Leonid Karlinsky, Tomer Michaeli and Ko Nishino (eds.). Springer Nature Switzerland, Cham, 104–120. https://doi.org/10.1007/978-3-031-25075-0_8

[63] Simone Stumpf, Evdoxia Taka, Yuri Nakao, Lin Luo, Ryosuke Sonoda, and Takuya Yokota. 2024. The Need for User-centred Assessment of AI Fairness and Correctness. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, June 27, 2024. ACM, Cagliari Italy, 523–527. https://doi.org/10.1145/3631700.3664912

[64] Chris Sweeney and Maryam Najafian. 2019. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019. Association for Computational Linguistics, Florence, Italy, 1662–1667. https://doi.org/10.18653/v1/P19-1162

[65] Synthesia. 2022. 60+ Languages | Different Voices & Accents | Synthesia. Retrieved December 20, 2022 from https://www.synthesia.io/features/languages

[66] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. 2021. Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. 1–16.

[67] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters* 140, (December 2020), 332–338. https://doi.org/10.1016/j.patrec.2020.11.007

[68] Lucas Whittaker, Kate Letheren, and Rory Mulcahy. 2021. The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing. *Australasian Marketing Journal* 29, 3 (August 2021), 204–214. https://doi.org/10.1177/1839334921999479

[69] Amber Grace Young, Ann Majchrzak, and Gerald C. Kane. 2021. Organizing workers and machine learning tools for a less oppressive workplace. *International Journal of Information Management* 59, (August 2021), 102353. https://doi.org/10.1016/j.ijinfomgt.2021.102353

[70] Yulei (Gavin) Zhang, Yan (Mandy) Dang, Susan A. Brown, and Hsinchun Chen. 2017. Investigating the impacts of avatar gender, avatar age, and region theme on avatar physical activity in the virtual world. *Computers in Human Behavior* 68, (March 2017), 378–387. https://doi.org/10.1016/j.chb.2016.11.052

[71] Sasa Zhao, Yanhui Xiang, Jiushu Xie, Yanyan Ye, Tianfeng Li, and Lei Mo. 2017. The Positivity Bias Phenomenon in Face Perception Given Different Information on Ability. *Front. Psychol.* 8, (April 2017), 570. https://doi.org/10.3389/fpsyg.2017.00570

[72] Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. 2024. Bias in Generative AI. https://doi.org/10.48550/ARXIV.2403.02726