Demographic Feature Isolation for Bias Research using Deepfakes

Kurtis G. Haut
khaut@u.rochester.edu
University of Rochester Department of Computer Science

Caleb Wohn
cwohn@u.rochester.edu
University of Rochester

Victor Antony
vantony@u.rochester.edu
University of Rochester

Aidan Goldfarb
agoldfa7@u.rochester.edu
University of Rochester

Melissa Welsh
mnwelsh13@gmail.com
University of Rochester

Dillanie Sumanthiran
dsumanth@u.rochester.edu
University of Rochester

M. Rafayet Ali
rafayet079@gmail.com
University of Rochester Department of Computer Science

Ehsan Hoque
mehoque@cs.rochester.edu
University of Rochester Department of Computer Science

ABSTRACT

This paper explores the complexity of what constitutes the demographic features of race and how race is perceived. “Race” is composed of a variety of factors including skin tone, facial features, and accent. Isolating these interrelated race features is a difficult problem and failure to do so properly can easily invite confounding factors. Here we propose a novel method to isolate features of race by using AI-based technology and measure the impact these modifications have on an outcome variable of interest; i.e., perceived credibility. We used videos from a deception dataset for which the ground-truth is known and create three conditions: 1) a Black vs White CycleGAN image condition; 2) an original vs deepfake video condition; 3) an original vs deepfake still frame condition. We crowd-sourced 1736 responses to measure how credibility was influenced by changing the perceived race. We found that it is possible to alter perceived race through modifying demographically visual features alone. However, we did not find any statistically significant differences for credibility across our experiments based on these changes. Our findings help quantify intuitions from prior research that the relationship between racial perception and credibility is more complex than visual features alone. Our presented deepfake framework could be incorporated to precisely measure the impact of a wider range of demographic features (such as gender or age) due to the fine-grained isolation and control that was previously impossible in a lab setting.

CCS CONCEPTS

• **General and reference** → **Cross-computing tools and techniques**; • **Applied computing** → *Law, social and behavioral sciences.*

MM ’22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3549204

KEYWORDS

Racial Bias, Confounding Factors, Artificial Intelligence, Credibility

ACM Reference Format:
Kurtis G. Haut, Caleb Wohn, Victor Antony, Aidan Goldfarb, Melissa Welsh, Dillanie Sumanthiran, M. Rafayet Ali, and Ehsan Hoque. 2022. Demographic Feature Isolation for Bias Research using Deepfakes. In *Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3503161.3549204

1 INTRODUCTION

Manipulating race as a variable to measure its impact has been extensively studied in the past. Bertrand and Mullainathan showed that having a stereo-typically Black name on a resume puts you at a disadvantage for employment opportunities compared to having a stereo-typically White name on an identical resume [6]. Avatars offer a high degree of customization and have been utilized to research impacts of race in virtual reality [12] and creating racial empathy that extends to the real world [26]. Although names and avatars are both effective ways to create racial perceptions, they cannot be applied to every aspect of racial bias. For example, while Bertrand and Mullainathan are able to measure call-back rates using names on resumes, they can’t study bias in a job interviews using the same technique (since racial perceptions in an interview would typically be based more on physical appearance rather than the name). The natural approach to study this topic would be to run an experiment where participants are shown one of two videos of mock job interviews where everything is identical (e.g., questions and answers) except the interviewee in one video is played by a White actor, and in the other is played by a Black actor. However, these actors will inevitably vary in more than just racial appearance, introducing confounding variables.

The very concept of race as a complex interplay of factors makes confounding variables seemingly inevitable. As Sen and Wasow argue, race is an aggregation of many elements including skin color, dialect, class, social status, and a myriad of other factors [32]. In order to measure racial bias, researchers must manipulate a subset of these elements. Doing so without introducing confounding variables, however, is easier said than done. A great deal of effort

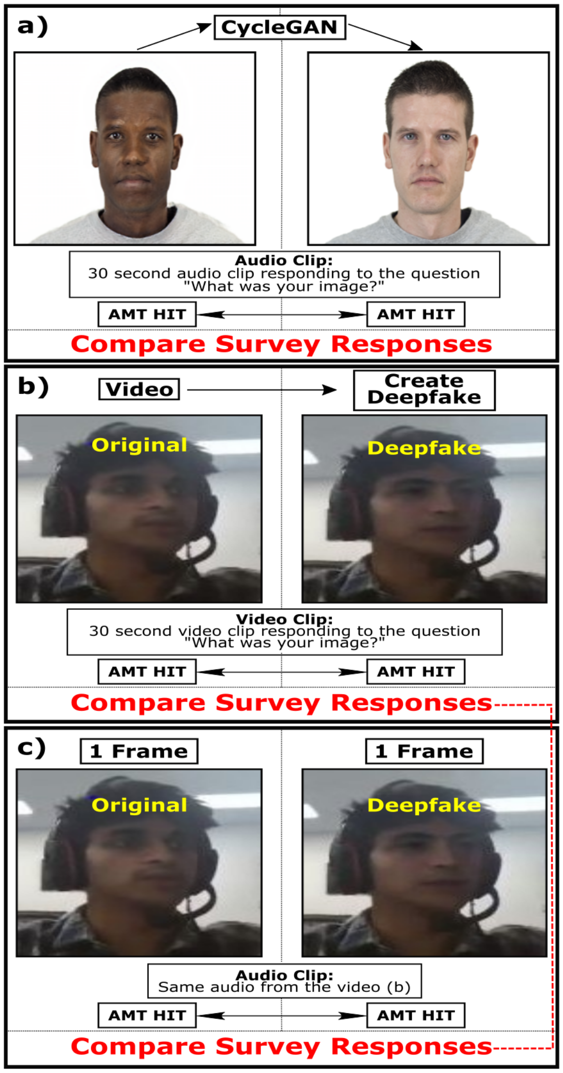


Figure 1: a) Black vs White CycleGAN image paired with same audio. We use CycleGAN [38] to racially map the Black image to a White image b) Original vs Deepfake video. We take an original video from UR Lying[33] deception database and create the deepfake with deepfacelab [27] using a White participants face c) Original vs Deepfake still frame paired with audio of b) We launched 6 surveys (2 for each condition) on Amazon Mechanical Turk (AMT HIT) and compared the responses to see if credibility changed due to the visual racial alterations seen in conditions a, b, and c.

has been directed at controlling for confounding variables in order to isolate the effects of race[15][36][29][19][9].

Recent advancements in AI allow us to control these confounding factors through the precise manipulation of skin tone, hair type, eye color, and facial features. This opens up an interesting opportunity to probe racial bias and measure it in a way that wasn't possible before. In this paper, we alter racial perceptions by leveraging deepfakes and a Cycle Generative Adversarial Network (CycleGAN) architecture to modify facial features and skin color while holding other variables such as accent and clothing constant. This allows us to directly measure the effect of those demographically visual racial features.

We apply our techniques on the publicly available UR Lying dataset, a naturalistic deception dataset collected using the Automatic Dyadic Data Recorder (ADDR) framework where the participants engage in an activity on a video call [33]. Credibility is important in many facets of life such as sales, negotiations, job interviews, medical appointments and criminal cases. Even for everyday human interactions, credibility assessments of our conversational partners impact communication. Therefore, understanding how racial perceptions influence credibility is worth exploring. We designed an experiment consisting of six surveys (see Fig. 1) to begin the process of answering this research question:

- How do racial perceptions in videos and images influence an individual's credibility?

We employed Amazon Mechanical Turk (AMT) to crowd-source responses for our surveys as Human Information Tasks (HITS, shown in Fig. 1). The participants are asked whether they believed the speaker was telling the truth, and we take credibility to mean the percent of participants who said yes. We have 6 surveys with 1736 total responses from the participants (For demographic information see supplemental materials). Each participant completed only one of the six surveys. The first two surveys test the effect of a CycleGAN mapping Black images to White images: The first survey measures credibility when the participant listens to audio from the UR Lying dataset while shown an image of a Black man (who they are told is the speaker). The second survey has the same audio, but the image is the output of the CycleGAN model on the image from the first survey – that is, when the speaker is made to appear White (see Fig. 1a). The next two surveys test the condition of altering race in a video. In one survey, a video from the UR Lying dataset is shown. In the other survey, we show a deepfake version of the same video where the speaker is made to appear more White (see Fig. 1b). In the final two surveys, we take one still frame from each of the videos (see Fig. 1c) while playing the audio from the video to see if there is a difference between video and image paired with audio for appearing credible. We used the CycleGAN [38] to generate image mappings of Black to White to test the Black vs White CycleGAN image condition. For original vs deepfake video condition, we used an encoder-decoder architecture to perform a face swap; essentially changing the visual representation of a darker skinned South Asian person to appear more White (in terms of skin tone and facial features) using DeepFaceLab [27]. By taking one frame from the original video and one frame from the deepfake, we also test an original vs deepfake still frame condition. The total number of responses for the six surveys is 1736. We performed statistical analysis on the survey responses to determine if the credibility differences for each condition were meaningful.

In summary:

- It is possible to alter racial perceptions through exclusively modifying visual representation of images and videos.

- We did not find the impacts to credibility due to these racial manipulations to be statistically significant.
- Our contribution is proposing a novel framework to isolate demographic features of race while holding the other features constant. Taking this approach, we can more precisely quantify how that feature is contributing to the observed bias.
- We note that this methodology can be extended to isolate other demographic features (such as gender or age) opening up new opportunities for bias research in variable domains.

2 BACKGROUND

2.1 Racial Bias

Racial bias permeates many aspects of society. Past research has found strong evidence that the criminal justice system is prone to racial bias [18, 28]. Black male offenders were even found to receive 19.1% longer prison sentences compared to White males, according to the United States Sentencing Commission in 2014 [4]. This racial bias extends to our school systems, with Black students receiving harsher punishments [35], creating a school-to-prison pipeline [22]. Racial bias also exists in our healthcare systems [11, 21], causing Black patients to receive lower quality care.

When assessing racial discrimination in context of credibility, prior works illustrate how people of different races tend to distrust each other the most [34]. The level of racial discrimination can even be broken down further into the skin color gradient [20]. For example, even when race is held as a constant and only skin color fluctuates, dark-skinned Blacks were 11 times more likely to experience racial discrimination than light-skinned Blacks [17]. These studies suggest that skin color alone may cause a person to experience bias. Günaydin et al. demonstrated that objective facial resemblance to a significant other influences snap judgments of liking automatically, effortlessly, and without conscious awareness [13]. Given that large proportions of individuals grow up in communities segregated or dominated by one racial group, Günaydin et al.'s research highlights the potential role racial perceptions play in credibility assessments. Understanding this relationship is important to morally explore to create fair, equitable societies. Without being perceived as credible, it is nearly impossible to convince the judiciary officials of your innocence, establish trusting relationships with your healthcare providers, or even receive the basic human respect to which all people are entitled.

2.2 Problem of Confounding Factors

In prior research, it has been difficult to change a visual feature like skin color without using different people. This makes it hard to isolate the visual features in question and inevitably leads to confounding factors. Even straightforward names as used in [6] cannot escape this problem. For example, if the name used is also associated with another factor, that factor could be the source of the measured bias and not race. Indeed, this excludability assumption must hold, necessitating further, meticulous studies in order to be able to draw valid conclusions concerning racial bias [2]. If even a variable as seemingly simple as a name is plagued by these underlying dependencies, then certainly visual racial features run into an even more complex web of entanglements. Yet despite these difficulties, we must control for these confounding factors and prior research involving race has worked hard to do so. A smoking study [5] was able to decouple racial segregation and socio-economic status and similarly, [36] controlled for socioeconomic status when researching the effect of race using regression techniques. However, [36] also makes clear that other factors such as skin color are much harder to control for. There are also statistical approaches to address confounding factors if they can be identified and their influence on the outcome variable can be reasonably estimated [1, 3]. Our research is able to control the effect of confounding factors because we can modify visual representation algorithmically on the computer. We can thus examine the effect a specific racial feature has on an outcome variable by altering that feature while holding every other feature constant. This paper chooses credibility as our outcome variable, but this framework would work equally well with different outcome variables and could be applied to features other than skin tone.

2.3 AI Techniques for Investigating Racial Bias in Credibility

CycleGANs have previously been used within research applications to make image processing training sets more inclusive to various skin tones. In the work of "Fairness GAN," Sattigeri illustrates the CycleGAN's power in constructing an extension to the Celebfaces Attributes to be demographically inclusive, one portion being skin tone [31]. Deepfakes can be applied to manipulate skin tone and facial features of videos and these algorithms continue to improve, generating more realistic images at a startling pace. In 2019, Karras et al introduced StyleGAN which demonstrated an innovative architecture able to generate more realistic images [16]. In order to more objectively assess how skin color and facial features effect one's credibility, we used techniques such as CycleGANs and deepfakes. The advantage these techniques offer is keeping features such as accent, clothing, gestures, and facial expressions constant while manipulating the variable of interest; namely visual racial features (see Fig. 1).

3 METHODS

Credibility Ground Truths - We used one audio clip and one video clip from the publicly available UR Lying dataset, collected using the ADDR framework [33]. Each clip is 30 seconds in length and encompasses the speaker answering the question "What was your image?". The speakers in the video are shown an image prior to answering this question. The ADDR framework instructed the speaker to lie or tell the truth about their image. In this particular instance, the speakers in all conditions described their image as it is (i.e., telling the truth). From the audio and video recordings, we designed six separate surveys to test the three conditions. The questions asked in each of the surveys remain consistent and we used Amazon Mechanical Turk (AMT) as the crowd-sourcing mechanism to collect responses. We received 1736 responses from unique users.

3.1 Black vs White CycleGAN Image Condition

We were interested to see how still images of people from different races paired with the same audio can influence credibility. The still image was designed to elicit a specific racial perception. We were interested in observing how credibility changes when a Black person is made to appear White. In order to generate the White image, we trained a CycleGAN using 4000 images from the Chicago Face Dataset (CFD) which learned how to conceivably change complex racial features such as eye color, lip shape, hair type and skin color [38]. The participants either saw the White image or the Black image, while listening to the same audio clip (see Fig. 1a). We compared the responses from these two surveys to see if the still image shown to the participants influenced the speaker's credibility.

3.2 Original vs Deepfake Video Condition

In the original vs deepfake video condition, participants either watched the original video in one survey or a deepfake video in the other survey. In the deepfake video, the person from the original video is made to appear more White using a DeepFaceLab face mapping [27]. While the CycleGAN model was able to generate impressive images, we wanted to study the effect of changing race in videos and CycleGAN was not appropriate for that task. Therefore, we used DeepFaceLab to generate the White video by taking the face of a White participant from the UR Lying database and mapped it onto the face of a darker skinned South Asian participant as seen in Fig. 1. Algorithmically, this is a faceswapping procedure initialized with src and dst video files. The dst video file is separated into an individual frame sequence based on its frame rate. A face extraction algorithm is then applied to both src and dst video files. Afterwards, these faces are aligned into pairs based off shortest distance from src and dst facial landmarks. These face pairs are used to train an autoencoder. The original frame sequence from the dst video file is then fed face by face through the trained model where the frame order is preserved. These frames (dst image with src image face) are then joined back together and combined with the audio from the dst video. The result is a deepfake video. Making a convincing deepfake video is not an easy task. In particular, making a dramatic change from Black to White is extremely difficult, and so we used a South Asian participant rather than a Black participant. In future work, other deepfake techniques should be investigated to see if this challenge can be overcome. We then create two surveys where one survey shows the participant the original video and the other shows the deepfake version of the video. We then compare the responses on the credibility of the speaker.

While the CycleGAN image condition created clear, unambiguous representations of race, the deepfake video condition explores the nuances of racial perception by making subtle modifications to the facial region.

3.3 Original vs Deepfake Still Frame Condition

We could not compare the first two conditions directly because the speakers are different people. Nonetheless, we were interested to see if there was a difference between a racial perception created via video versus a still image. Whether this has an effect on credibility could be important for video calling environments where some people use video while others just have a still image of themselves. In order to directly compare, we first take one still frame each from the original video and the deepfake video respectively (see 1c). Similar to the other two conditions, we then launch two separate surveys. We pair these still frames with the audio from the video.

3.4 Assessing Credibility Differences From Changing Perceived Race

In this experiment, we measure credibility by asking the participants whether they thought the speaker was lying or telling the truth and the credibility of the speaker is simply the percentage of participants who believed the speaker was telling the truth. To assess differences in credibility, we took the credibility percentage in each condition and compared them against each other using a proportions Z-test. We also asked the participants what race they thought the speaker was in each survey to quantify racial perception. Since we have metrics on perceived race and credibility, and the only feature changing across conditions are visual indicators of race, we can start to explore how race might influence credibility.

4 RESULTS

We found that our method was able to change race perception. Fig 2 shows that the perception of race in the CycleGAN imgage condition changed from predominantly Black to White. Similarly, fig 3 shows that the deepfake video and still frame were also able to alter race perception. However, we did not find statistically significant differences in credibility in any condition (Table 1 shows these results).

Table 1: Credibility results from all six surveys

Survey	n	Credibility
Black Image	149	70.47%
White CycleGAN Image	162	71.60%
Original Video	437	70.25%
Deepfake Video	467	66.81%
Original Still Frame	274	71.53%
Deepfake Still Frame	247	68.42%

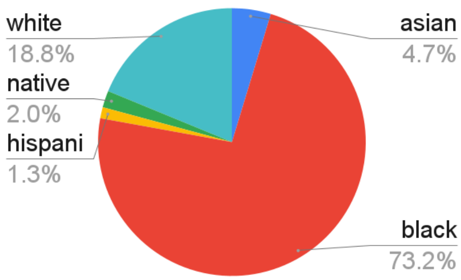
5 DISCUSSION

5.1 No Credibility Differences

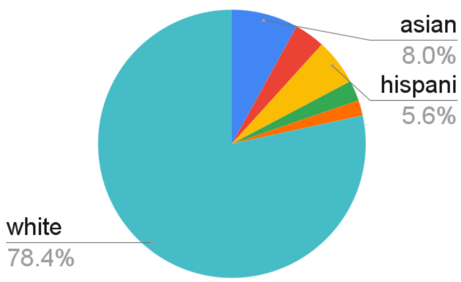
Given the extensive evidence of racial bias found by prior research, we were surprised to find that making a speaker appear White (or more White) did not impact their credibility. Perhaps the reason that our results do not detect racial bias in credibility is that we are only altering facial features and skin tone. Other features of race could have a larger impact on credibility and, by using our methodology, we can measure that effect by considering the features in isolation. Another possible explanation could be that our techniques to alter racial perception introduce small flaws into the video which make the video feel less authentic, arousing suspicion (more advanced deepfake technology may be able to solve this issue).

5.2 Technical Challenge of Altering Racial Perceptions Using Deepfakes

For example, we tried to train the CycleGAN using images of White and Black participants from the UR Lying dataset. Unfortunately,



a) Black Image Survey

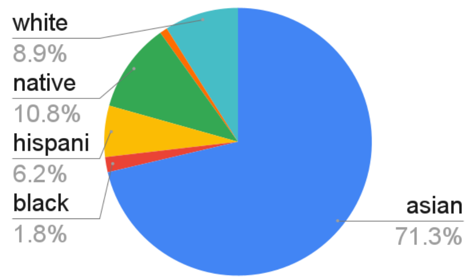


b) White CycleGAN Image Survey

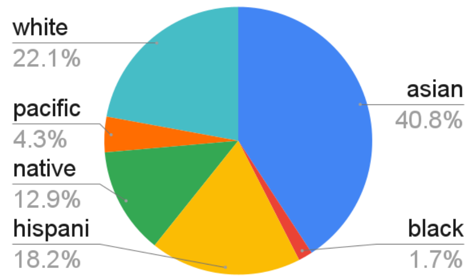
Figure 2: Responses to "What race do you think the speaker is?

the participants were wearing headsets (see Fig. 1b,c) and their backgrounds were inconsistent. We got around this issue by utilizing Chicago Face Dataset (CFD) where the images were high resolution head shots with consistent clothing and backgrounds across the dataset (see Fig. 1a). However, this did not solve the problem of creating a realistic deepfake video. If we were able to successfully train the CycleGAN on images from the UR Lying dataset, we could have used it to map images in sequence to change someones race from Black to White. Then, we could overlay the desired audio (Assuming the mapping was accurate enough to preserve normal lip sinc). We shopped around and went with deepfacelab because it takes the problem of generating the scene out of play by performing a face swap. This posed another problem, in the process of mapping a White person's face onto a Black persons body, the result is a comically unconvincing person with a White face, Black neck and, from our dataset, a distinguishing African hair type. Thus, we were forced to use a darker-skinned South Asian person who came with the added benefit hair ambiguity (see Fig. 1b). Even then, many manual hours were required using the image processing/enhancement (face extraction, blur, glitch) toolkit provided by DeepFaceLab to make believeable deepfake.

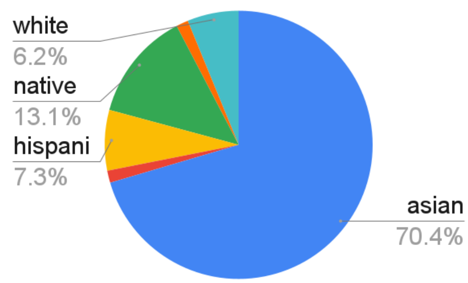
In addressing these technical challenges, We found that racial perception can be changed by modifying visual representation exclusively. This can be seen most clearly through our Black vs White CycleGAN image condition (see Fig. 2). We were able to invert the racial perception from Black to White using our CycleGAN image mappings. Through our original vs deepfake video condition and still frame condition, we see that the landscape of race can also be modified in a more nuanced way. This can be observed



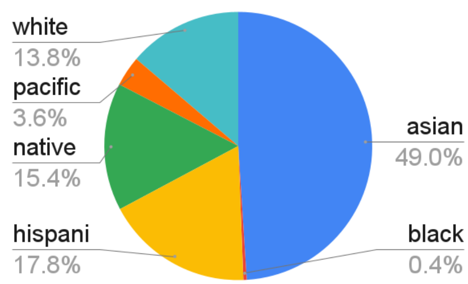
a) Original Video



b) Deepfake Video



c) Original Still Frame



d) Deepfake Still Frame

Figure 3: Responses to "What race do you think the speaker is?

by looking at the distribution of race responses in Fig. 3. We see from this distribution that image and video are both able to alter race perception and there is relatively no difference between them. Future work could probe and compare different racial filters using more advanced deepfake technology. More work is needed to also explore the bi-directional alternation of races to understand the whole spectrum of racial perception.

5.3 Limitations

Issues of Video Realism - To address this potentially confounding factor in our study, we posed the question "I found the video to be authentic" in most of the original vs deepfake video surveys. For the original video, 291/328 (88.1%) of participants found the video to be authentic while for the deepfake video, 308/371 (83.3%) of participants thought the video was authentic. After running a proportions z test we concluded that there is no statistically significant difference between these surveys regarding the perceived authenticity of the videos.

Collecting Data from Amazon Mechanical Turk - We use AMT as a mechanism to draw a sample of the general population. We must recognize that this surveying technique will suffer from some degree of population bias. Unfortunately, the cost of bringing people into the lab to complete the surveys was made burdensome due to the social distancing stipulations of the pandemic. It also would have been far more expensive and would implement a population bias of a different kind. In fact, generally, mTurkers form a more diverse population than the standard internet population or the populations of college students used in laboratory experiments [8][14][25]. Considering the trade-offs, AMT was our best option given its experimental validity and ability to crowd-source a large number of responses effectively. To see the demographics of the mTurk population, please see the supplementary materials.

5.4 A Framework for Demographic Feature Isolation in Bias Research

Our major contribution is proposing a systematic feature isolation framework that can be applied to research bias by controlling for complex confounding factors. For instance, how race is perceived depends on many other factors besides solely visual representation. Dialect, social class, geographic location, and plenty of others all contribute to the complexity of race perception and any one, or combination of these individual factors could cause racial bias. To properly understand the extent each racial feature is contributing to the bias, the entire feature space needs to be controllable and finetunable. Then, by manipulating a single feature while the rest remain unchanged, we can measure the impact that one feature has. We measure the features impact towards the racial bias by having an outcome variable of interest. Having a way to adequately and consistently measure this outcome variable is equally important because that is how the bias is approximated. In our case, our outcome variable is credibility and we measured it as the percentage of survey respondents who believed the speaker was telling the truth. When both of these criteria e.g., controlling the feature space and measuring outcome variable appropriately are met, our framework can be used to control confounding factors and research different aspects of certain types of bias given an outcome variable of interest.

5.5 Future Work

This paper focuses on examining what happens when the visual representation of race is changed with AI and how that influences one's credibility. However, our approach of systematically isolating demographic features enables new research opportunities for studying a variety of biases (such as gender and age) and exploring their intersectionality could unearth fresh insights in the future. Apart from visual alternations, research could focus on isolating specific accents and comparing that against Standard American English to observe the effect this has on an individual's credibility. Furthermore, Our methodology can be applied to research bias in areas beyond just credibility, such as perceived intelligence, competency, or likeability. We believe that this methodology could be improved with more advanced deepfake techniques and applied to a variety of contexts to explore racial bias. In particular, lip-sync technology such as [30] and [37] show great promise for this application – the still images produced by the race-swapping CycleGAN model are very believable, and convincing lip-sync animations of these images would be perfect for this experiment (unfortunately, when we attempted to create such animations using the existing lip-sync models previously mentioned, the resulting videos were obviously fake)

5.6 Ethics of Altering Race with Deepfakes

We are aware of the sensitive nature of race as a topic of research, and feel it is necessary to discuss the ethical concerns surrounding our proposed framework in the realm of peer reviewed science to instigate a broader dialogue.

The salient concern with altering race perception with deepfakes is that it could reinforce racial stereotypes or be used in a demeaning way. In our controlled experiment, we believe that these harms are minimal and are outweighed by the potential benefits this approach offers for research into racial bias.

There is a large body of literature on the dangers of deepfake technology. Deepfakes pose a variety of risks such as impersonation, deception, and fake news [7, 24]. We must keep these concerns in mind when applying deepfakes. However, we think that our research represents a case where deepfakes can be used in a positive way to produce insights that would not be derived otherwise. A more robust understanding of racial bias could be leveraged to help create a more fair and equitable society.

In an effort to promote fairness and justice, it is increasingly important to consider legal and ethical problems arising from potential bias in AI models [23]. [10] outlines the issue of unavoidable biases that exist in visual datasets. In our case, the Chicago Face Dataset could have a favorable credibility bias due to their minimalistic head shot representation and professional photo quality. As a result, there are fewer idiosyncratic features that could subconsciously provoke distrust (such as tatoos or piercings). This could cause the higher credibility ratings which we observed. The models we utilized to generate the deepfakes also come with concerns of algorithmic bias. We speculate that CycleGAN is biased towards images that have underwent preprocessing to control for unique image elements. Its possible that this preprocessing could introduce unintended bias if individual uniqueness is phased out during training (recall, CycleGAN could not adapt to the headsets worn by subjects the UR Lying dataset). The encoder-decoder architecture used by deepfacelab may suffer from algorithmic bias. However, the strongest sources of bias are likely the result of decisions made by users of the software rather than default model training parameters.

Another problem which gravely concerns us is the risk of our results being misinterpreted or overstated. Although we did not find significant differences as a result of our manipulations, this should *not* be used to argue against the well-established fact of racial bias.

6 CONCLUSION

We began the process of quantifying how changing demographic features of race influences credibility. This is a difficult problem, and relatively understudied, as isolating the variables contributing to racial profiling is challenging. Here we isolated one of those variables – visual representation (e.g., skin tone and facial features) – and explored how modifying visual representation impacts credibility. We used AI to assist in creating and modifying specific race perceptions. CycleGAN was used to create a White representation of a Black image, while deepfakes were used to modify race in a video. We created surveys to test the various image and video conditions using 1736 online workers from Amazon Mechanical Turk. By comparing the responses for the surveys associated with each condition, we measure the effect of how altering demographically visual features changes racial perception and influences credibility. While our data shows no significant differences in credibility across the conditions, it is important to probe this matter further. We propose our framework for isolating demographic features and researching different aspects of bias. Our framework shows a way to study something as complex as race and credibility in a novel fashion and opens the door for additional bias research involving gender or age. Using advances in AI, we can begin answering questions that we couldn't study before. We hope this work serves as an initial investigation into this space and encourages further exploration.

ACKNOWLEDGMENTS

This work was supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under grant W911NF19-1-0029.

REFERENCES

[1] [n.d.]. National Library of Medicine How to control confounding effects by statistical analysis. How to control confounding effects by statistical analysis. Accessed: 2022-04-14.

[2] [n.d.]. Political Analytics An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments. www.cambridge.org/core/journals/political-analysis/article/an-empirical-justification-for-the-use-of-racially-distinctive-names-to-signal-race-in-experiments/DBC39F875F2DC0F65E7140FC721CE1EB. Accessed: 2022-04-14.

[3] [n.d.]. Science Direct Confounding: What it is and how to deal with it. https://www.sciencedirect.com/science/article/pii/S0085253815529748. Accessed: 2022-04-14.

[4] [n.d.]. United States Sentencing Commission demographic sentencing. https://www.ussc.gov/research/research-reports/demographic-differences-sentencing. Accessed: 2022-04-14.

[5] [n.d.]. Wiley Online Library Overcoming confounding of race with socio-economic status and segregation to explore race disparities in smoking. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1360-0443.2007.01956.x. Accessed: 2022-04-14.

[6] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[7] Johnny Botha and Heloise Pieterse. 2020. Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*. 57.

[8] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? (2016).

[9] Daniel M Butler and Jonathan Homola. 2017. An empirical justification for the use of racially distinctive names to signal race in experiments. *Political Analysis* 25, 1 (2017), 122–130.

[10] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. 2021. A survey on bias in visual datasets. *arXiv preprint arXiv:2107.07919* (2021).

[11] Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics* 18, 1 (2017), 1–18.

[12] Victoria Groom, Jeremy N Bailenson, and Clifford Nass. 2009. The influence of racial embodiment on racial bias in immersive virtual environments. *Social Influence* 4, 3 (2009), 231–248.

[13] Gül Günaydin, Vivian Zayas, Emre Selcuk, and Cindy Hazan. 2012. I like you but I don't know why: Objective facial resemblance to significant others influences snap judgments. *Journal of Experimental Social Psychology* 48, 1 (2012), 350–353.

[14] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental economics* 14, 3 (2011), 399–425.

[15] KJ Jager, C Zoccali, A Macleod, and FW Dekker. 2008. Confounding: what it is and how to deal with it. *Kidney international* 73, 3 (2008), 256–260.

[16] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[17] Elizabeth A Klonoff and Hope Landrine. 2000. Is skin color a marker for racial discrimination? Explaining the skin color–hypertension relationship. *Journal of behavioral medicine* 23, 4 (2000), 329–338.

[18] Margaret Bull Kovera. 2019. Racial disparities in the criminal justice system: Prevalence, causes, and a search for solutions. *Journal of Social Issues* 75, 4 (2019), 1139–1164.

[19] Thomas A LaVeist, Roland J Thorpe Jr, GiShawn A Mance, and John Jackson. 2007. Overcoming confounding of race with socio-economic status and segregation to explore race disparities in smoking. *Addiction* 102 (2007), 65–70.

[20] Zeus Leonardo. 2004. The color of supremacy: Beyond the discourse of 'white privilege'. *Educational philosophy and theory* 36, 2 (2004), 137–152.

[21] Ivy W Maina, Tanisha D Belton, Sara Ginzberg, Ajit Singh, and Tiffani J Johnson. 2018. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social Science & Medicine* 199 (2018), 219–229.

[22] Jason P Nance. 2015. Over-disciplining students, racial bias, and the school-to-prison pipeline. *U. Rich. L. Rev.* 50 (2015), 1063.

[23] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[24] Konstantin A Pantserev. 2020. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In *Cyber defence in the age of AI, smart societies and augmented humanity*. Springer, 37–55.

[25] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.

[26] Tabitha C Peck, Sofia Seinfeld, Salvatore M Aglioti, and Mel Slater. 2013. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and cognition* 22, 3 (2013), 779–787.

[27] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework. *ArXiV* (2020).

[28] Joan Petersilia. 1985. Racial disparities in the criminal justice system: A summary. *Crime & Delinquency* 31, 1 (1985), 15–34.

[29] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. 2012. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench* 5, 2 (2012), 79.

[30] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*. 484–492.

[31] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. 2018. Fairness gan. *arXiv preprint arXiv:1805.09910* (2018).

[32] Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science* 19 (2016), 499–522.

[33] Taylan Sen, Md Kamrul Hasan, Zach Teicher, and Mohammed Ehsan Hoque. 2018. Automated dyadic data recorder (ADDR) framework and analysis of facial cues in deceptive communication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–22.

[34] Sandra Susan Smith. 2010. Race and trust. *Annual Review of Sociology* 36 (2010), 453–475.

[35] Cheryl Staats. 2014. Implicit racial bias and school discipline disparities. *Exploring the connection* (2014).

[36] Tyler J VanderWeele and Whitney R Robinson. 2014. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)* 25, 4 (2014), 473.

[37] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.