Exploring the Racial Bias in Pain Detection with a Computer Vision Model

Sarah Fabi (sfabi@ucsd.edu)
Department of Cognitive Science and Halicioglu Data Science Institute, University of California San Diego, USA and
Neuro-Cognitive Modeling Group, University of Tübingen, Germany

Xiaojing Xu (xix068@ucsd.edu)
Department of Electrical and Computer Engineering, University of California San Diego, USA

Virginia R. de Sa (desa@ucsd.edu)
Department of Cognitive Science and Halicioglu Data Science Institute, University of California San Diego, USA

Abstract

People detect painful expressions more easily in members of their racial ingroup than outgroup. Here, we wanted to investigate this racial bias with a machine learning model trained to detect activations of different action units of painful facial expressions. We examined whether the model detected higher action unit activation for European than African faces when trained on datasets with mostly White faces. To control for confounding variables, pictures of faces were generated with the FaceGen Modeller. Results revealed that there exist differences in the visual detectability of some facial muscle activations due to skin color or other race-dependent facial features. Despite the bias towards European looking faces in the training data, some activations were more easily detectable in African faces. Thus, neither the perceptual detectability, nor the larger exposure to own-race faces seems to solely explain the racial bias in pain detection.

Keywords: racial bias; pain recognition; FACS; automated pain recognition; action units; facial expression recognition

Introduction

Pain is often underestimated in people of color, leading to race-based physical health disparities (Kissi, Van Ryckeghem, Mende-Siedlecki, Hirsh, & Vervoort, 2022; Mays, Cochran, & Barnes, 2007). When experiencing others in pain, people show more behavioral and neural empathic responses if the others are members of their racial ingroup compared to members of an outgroup (Fabi & Leuthold, 2018; Sessa, Meconi, Castelli, & Dell'Acqua, 2014; Xu, Zuo, Wang, & Han, 2009). Understanding this racial bias in empathy for pain and finding preventive measures is crucial to prevent the aforementioned outcomes. The aim of this work is to investigate such differences in pain perception using a machine learning model. We were interested in the question of whether there exist differences in the visual detectability of some facial muscle activations that can explain the difficulties in detecting pain in outgroup members. Therefore, a ML model, which is trained on detecting facial action units associated with painful expressions rated the activation of facial action units from generated faces. The stimuli were pictures of artificially generated avatar faces with the software FaceGen Modeller that allows to not only keep everything constant except for one feature (e.g. skin color) but also to manipulate the different action units individually. With this, we can investigate whether the computer vision (CV) model is judging the same faces differently if they vary a) just in skin color or b) in racial features like shape of eyes, nose etc. and make more detailed tests for specific action units.

Theoretical Background

Facial expressions allow probabilistic inference of emotional state, for instance how much pain, surprise, disgust, etc. a particular person is experiencing. In order to classify facial expressions, researchers developed a comprehensive as well as psychometrically rigorous taxonomy for muscle movements called the "Facial Action Coding System" (FACS) (Hjortsjö, 1970; Cohen, Ambadar, & Ekman, 2007). This system, which was updated three times between 1978 and 2002, defines a variety of "Action Units" (AUs), which can be identified by decomposing facial expressions into the smallest discriminable facial movements. Since nearly all humans share the same facial muscles (Schmidt & Cohn, 2001) and hence the same anatomical basis for facial movements, FACS associates the movement of one of more facial muscles with discrete AUs. The list of AUs starts with the upper face, meaning brow actions, eye region actions, etc. Further AUs are located in the lower face, coding mouth and chin movements.

The FACS taxonomy in and of itself does not include any instructions on how to infer mental states. Here, other psychometric frameworks containing combination rules for AUs must be put to use or learned (Xu & de Sa, 2020). The Prkachin and Solomon Pain Intensity (PSPI) (Prkachin & Solomon, 2008) describes the amount of pain detected in a facial expression by calculating a sum of different AUs: $PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43$. All AUs range between 0 and 5, except for AU43, which can either be 0 or 1 (eyes open vs. closed), leading to a maximal PSPI value of 16. Apart from the PSPI, the following AUs are known to be non-exclusively related to painful facial expressions: AU4: brow lowering, AU6: cheek raising, AU7: eyelid tightening, AU9: nose wrinkling, AU10: upper lip raising, AU12: oblique lip raising, AU20: horizontal lip stretching, AU25: lips parting, AU26: jaw dropping, AU43: eye closing.

Until a few years ago, video-recorded facial expressions had to be coded manually in a labor-intensive process. But in light of recent progress in machine learning, facial recognition algorithms can be utilized to automatically classify AUs with promising results in accuracy. In this context, machine learning methods for automatic pain detection and estimation pose an increasingly important tool for smart health-

care applications. They are applied on humans (Xu et al., 2018; Tavakolian & Hadid, 2018, 2019; Kaltwang, Rudovic, & Pantic, 2012; Martinez, Rudovic, & Picard, 2017; Zhou, Hong, Su, & Zhao, 2016; Xin, Lin, Yang, & Zheng, 2020; Zafar & Khan, 2014; Soar, Bargshady, Zhou, & Whittaker, 2018; Bargshady et al., 2020a; Lopez-Martinez & Picard, 2018; Guo, Wang, Xiao, & Lin, 2021; Bargshady et al., 2020b; Tavakolian & Hadid, 2018; Tavakolian, Bordallo Lopez, & Liu, 2020; Thuseethan, Rajasegarar, & Yearwood, 2019; Xu & de Sa, 2021; Susam et al., 2021) as well as animals (Andresen et al., 2020; Lencioni, de Sousa, de Souza Sardinha, Corrêa, & Zanella, 2021; Noor et al., 2020). Not all of the approaches for automatic pain estimation use facial expressions as input (Susam et al., 2018; Pouromran, Radhakrishnan, & Kamarthi, 2021), but many do. Most of those acting on facial expressions use convolutional neural networks or, to be more precise, fine-tuned VGG-Face (Parkhi, Vedaldi, & Zisserman, 2015) since these models prove to be well-adjusted to feature identification in (facial) images.

In humans, the racial bias in empathy for pain might in part be attributable to difficulties in detecting the correct AU activations in painful faces belonging to the racial outgroup. Mende-Siedlecki, Qu-Lee, Backer, and Van Bavel (2019) investigated pain detection in White participants, when being confronted with painful facial expression of the racial in- and outgroup. Across seven experiments, they came to the conclusion that White participants show consistent difficulties in detecting the pain in Black faces, even if the exact same facial expression was shown on in- and outgroup targets. This also held true for people who did not have an explicit racial bias (assessed as the difference of feeling warmth for White and Black persons, embedded into eight other social groups).

There exist several competing explanations for the emergence of this racial bias in pain perception. Two of them shall be examined more closely in this work. First, the racial bias in pain perception could be due to perceptual properties in Black faces like the shape, color, or contrast of specific facial features which make it harder to detect the activation of pain-related AUs. Or, second, White subjects are more often exposed to facial expressions of their ingroup members, meaning that they have more data on which they can train their pain recognition skills in ingroup faces. With human subjects, it is hard to investigate these mechanisms in detail, since the subjects' previous exposure to members of different races is hard to measure. Therefore, we wanted to investigate these questions with one of the machine learning models that was specialised on detecting pain in faces (Xu, Huang, & de Sa, 2020) and for which we knew the training dataset and thus how much proportional exposure it had to different races. Here, we will not focus on biases in machine learning models (Barocas & Selbst, 2016) – which are not necessarily a shortcoming (Fabi & Hagendorff, 2022) – but on investigating the human racial bias in pain recognition with a machine learning model. Furthermore, with pictures of real persons in pain as stimuli, the perceptual properties cannot be examined in detail because of confounding variables like different manners to express pain, differences in natural faces etc.. Therefore, we chose to apply pictures of avatar faces, for which we could keep everything constant except for the features we wanted to investigate. Additionally, we were able to determine an objective label for the AU activations, since we manipulated the AUs ourselves.

Method

Computer Vision Model

We performed our experiments with the first stage of the CV model in (Xu et al., 2020) which was specifically trained to detect AUs related to pain, the PSPI values, and the values of the Visual Analog Scale (VAS). The whole model consisted of three stages: The first stage was a neural network trained to predict frame-level PSPI and AUs. The second stage was a fully connected neural network to predict sequence-level pain scores from the PSPI predictions of stage 1. The third stage combined these pain scores in an optimal linear manner to output a final VAS value. In more detail, the first stage is based on VGGFace (Parkhi et al., 2015) which was pre-trained to classify 2622 faces of famous individuals, of which the majority are White. Xu et al. (2020) replaced the last layer with a linear fully-connected regression layer and trained the whole network additionally on the UNBC-McMaster Shoulder Pain Dataset (Lucey, Cohn, Prkachin, Solomon, & Matthews, 2011) to detect the PSPI value for each frame of the videos, as well as the following nine AUs, which were present in at least 500 frames in the training dataset: AU 4, 6, 7, 10, 12, 20, 25, 26, 43. The AUs were originally predicted together with the PSPI to improve the PSPI prediction. In our case, we are mainly interested in the AUs. No training on our avatar dataset was done.

Stimuli

Most of the automatic pain detection research work draws upon the UNBC-McMaster Shoulder Pain Dataset (Lucey et al., 2011) for training and testing classification models. The dataset consists of 200 videos of 25 shoulder pain patients performing various active and passive range-of-motion tests to their affected and unaffected limbs. Subjects are mostly White, representing more or less a cross-section through the Canadian population. In our experiments, though, we do not utilize real-world, but synthetic images, that allows us to eliminate confounds. FaceGen Modeller is a tool to generate random faces that can be manipulated in various ways, including by race, gender, and age. Furthermore, each AU can be activated within a scale between 0 and 10, eliminating influences of human raters who do not always agree on AU activations (De la Torre, Simon, Ambadar, & Cohn, 2011).

The pictures of the avatar faces were generated in the FaceGen Modeller following a protocol. Per randomly generated male face (with random texture and features belonging randomly to different races), we created one light and one dark skin color version by setting the skin shade to 2.5 and -3.0,

(a) Dark and light skin color condition of a random face.

(b) African and European condition of the same face.

Figure 1: Example stimuli of the same randomly generated face in the four conditions.

respectively. Next, color was set to zero and the facial features and color were changed to more European- or African-looking. Then, painful expressions were created by manipulating the activation of the nine AUs the CV model can predict. The exact same facial expression was then applied to all four different conditions of the randomly generated face (cf. Figure 1). We applied this procedure to 25 different base faces, leading to 100 face pictures in total. For image preprocessing, we used the cascade DPM Face Detector (40; 41) to detect the face and then extended the bounding box by a factor of 0.1 when cropping the face. We then resized the image to 224×224 and normalized each channel with the mean and standard deviation of the data the model was pre-trained on.

Experiments

In the first experiment, we wanted to investigate whether the skin color change on random faces alone led to higher or lower predictions of AU activations. Therefore, we fed the CV model with the 50 faces of the light and dark skin shade condition and analysed the results with paired t-tests. In the second experiment, the CV model got the European and African faces as inputs to detect additional differences based on facial features. In both experiments we started with the sum over all AU activations and continued with tests for the nine different AUs with a Bonferroni corrected alpha of .0056. In follow-up tests, we looked into specific AUs of specific faces to determine the reason underlying the differences between the conditions.

Results

Across 25 faces, the sum of all predicted AU activations was not significantly different for the dark vs. light skin shade condition ($t = -0.62, p = .54$), nor between the European and African condition ($t = 0.12, p = .91$). This means that over all AUs, the CV model did not detect more activation for the dark than light skin color or for one of the two races. If all AUs were equally relevant for pain, the CV model would not detect more pain in one race than the other. But since different AUs vary in their importance for painful expressions, we wanted to look into specific AUs.

For the light versus dark skin shade conditions, we found larger values for dark skin in AU 6 ($p < .0001$) and 10 ($p < .0001$). The predictions were higher for the light skin color in AU 25, 26, and 43 (all ps $< .005$).

In line, predictions for AU 6 and 10 were higher in the African vs. European condition and 25 was higher in the European condition (all ps $< .0001$). AU 43 was not significantly different between races ($p = .12$). AU 26 was even higher for the African than the European condition ($p < .0001$). For the latter two AUs, the morphological feature effects seem to outweigh the difference of skin shade. Furthermore, AU 4, 7 and AU 20 were now significantly higher in the European than the African condition (all ps $< .0001$). In sum, the results of the tests for skin shade and racial features indicate that for AU 6, 10, and 25 the differences between races might be mostly due to skin shade and contrast, whereas the differences for AU 4, 7, 20, and 43 are not (solely) attributable to skin color. The racial features seem to play a bigger role here. Since our dark and light faces had not the exact same color as the African and European faces, respectively, these interpretations have to be taken with caution. Therefore and in order to investigate the specific racial features in more detail, we decided to look at some AUs in more detail.

Investigating specific AUs

Based on previous results of the CV model of Xu and de Sa (2021), we selected two AUs that the model could detect with high accuracy and for which the range of predicted values was high: AU 25 and 7. We intentionally selected AUs, which were better detected for the light or European condition to get insights regarding the racial bias in pain detection in White humans. The specific AUs were examined with the help of one avatar face of the stimuli set that had the specific AU maximally activated and showed strong differences between the two conditions. First, we wanted to investigate whether the AUs were correctly identified for our avatar faces, since the CV model had previously seen only real human faces. Therefore, we hold every AU in the specific face constant, except for the AU that should be examined: This was activated at 0, 25, 50, 75 or 100%. Such morphs were created for the African and European face. Second, we investigated the influence of the skin color on our results by giving the African face a European skin color and the European face an African skin color. If the skin color could not solely explain the differences between the two races, the African face with the European skin color was incrementally made more European looking in order to determine whether specific facial features were responsible for the differences in the detection of the AUs.

AU 25: Lips parting AU25 describes how far the lips are parted due to action of depressor labii, relaxation of mentalis, and orbicularis oris. As described above, over all 25 faces,

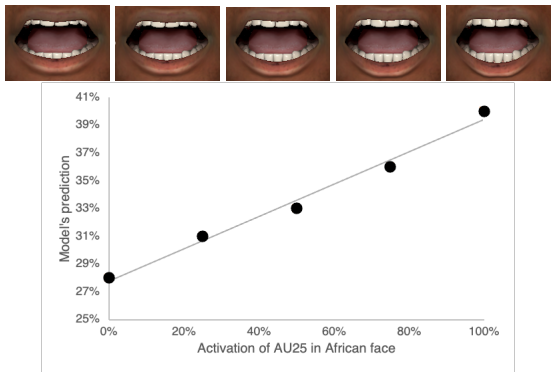


Figure 2: The CV model detects the increase in AU25 activation in African faces. The full faces were shown, but we show insets of just the mouth to better allow comparison for the reader.

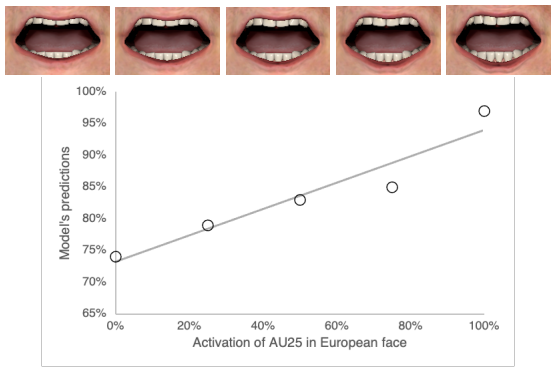


Figure 3: The CV model detects the increase in AU25 activation in European faces. The full faces were shown, but we show insets of just the mouth to better allow comparison for the reader.

the CV model detected higher AU25 activations for dark than light and for African than European faces. For the specific test, we show face 21 with an AU25 activation of 100% (see Figure 4a), though the results are consistent across other faces as well. The CV model detected AU25 activation of 97% for the European and of 40% for the African face. The model detected increasing AU activations for the morphs of both races, ranging between 74 and 97% and between 28 and 40% for the European and African faces, respectively. Even though the variance between the values is not very high, the results in Figures 2 and 3 show that the CV model did a very good job in detecting the relative AU activation. This is even more astonishing when looking at how subtle the differences of the AU25 activation were. The results allow the conclusion that FaceGen Modeller manipulated the AU activations in a reasonable manner.

Next, we investigated our hypothesis that most of the AU25 differences between the two races was due to the skin color since the differences of the dark and light skin shade stimuli was also significant. Since the dark and light faces had not the same skin color as the European and African faces, we

(a) CV model's prediction: 40 vs. 97% AU25 activation

(b) CV model's prediction: 103 vs. 32% AU25 activation

Figure 4: (a) African and European face with several AU activations, including AU25 at 100%. (b) Same faces with skin color of the opposite race. Skin color seems to explain most of the AU25 differences.

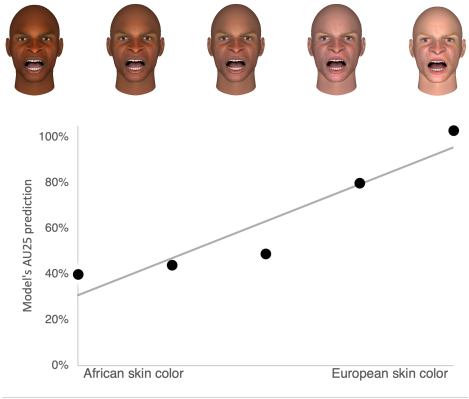


Figure 5: The CV model's AU25 prediction for the African face increases when changing the skin color.

now gave the European and African face the skin color of the opposite race (cf. Figure 4b). The face with African features but European skin color led to a detection of 103% AU25 activation, whereas the face with European features but African skin color led to a detection of 32%. These values are very similar to the original detection of 97 and 40%, supporting our hypothesis that skin color and in this case possibly the contrast between skin, lips, and teeth was responsible for the higher detection of AU25 in European vs. African faces. For a more incremental change of the African face's skin color, see Figure 5.

AU7: Eyelid tightening Next, we examined the lid tightener (AU7) which was more highly detected over the 25 European versus African faces. However, the CV model did not detect significant higher activation for our light than dark skin shade condition. This led to the hypothesis that, unlike AU25, the difference in AU7 was not solely due to the skin color. To investigate this hypothesis and examine the importance of dif-

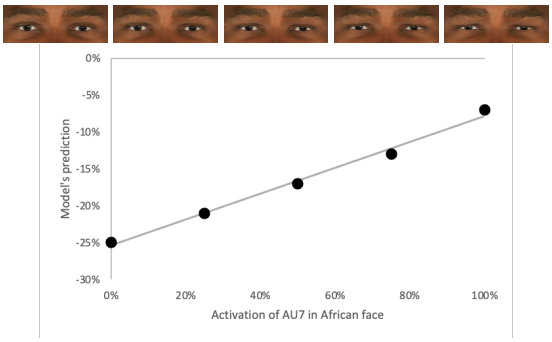


Figure 6: The CV model detects the increase in AU7 activation in African faces. The full faces were shown, but we depict insets of just the eye region to better allow comparison for the reader.

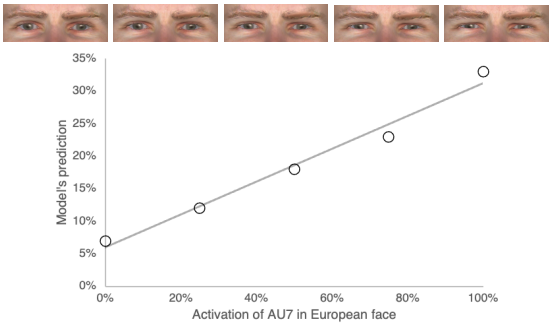
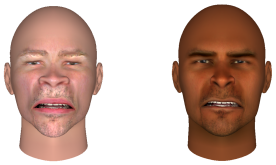


Figure 7: The CV model detects the increase in AU7 activation in European faces. The full faces were shown, but we depict insets of just the eye region to better allow comparison for the reader.

ferent racial features, we selected face 14, for which the CV model rated the AU7 activation (which was originally 100) as -7 and +33% for the African vs. European face, respectively (see Figure 8a). The results for this face reported below are not cherry-picked but hold true across different faces.

Again, we tested the overall ability of the CV model to detect the differences in AU7 activation in the morphs of one race. The model detected the increase in AU7 activation at every step for both races, ranging from 7 to 33% and from -25 to -7% for the European and African face, respectively (see Figures 6 and 7). The color test revealed that the CV model detected 15% of AU7 activation for the African face with European-like skin color and 17% for the European face with African-like skin color (cf. Figure 8b). This means that the light skin color led to slightly higher AU7 detection than the dark skin color in African faces, but the CV model's prediction was not close to the 33% for the light European face. Making the European face dark led to a decrease in the detection but it was still much higher than for the dark African face (-7%). This speaks to the fact that regarding the AU7 difference, skin color is important, but is not the only factor. For a more incremental change of the African face's skin color, see Figure 9.

(a) CV model's prediction: -7 vs. 33% AU7 activation



(b) CV model's prediction: 15 vs. 17% AU7 activation

Figure 8: (a) African and European face with several AU activations, including AU7 at 100%. (b) Same faces with skin color of the opposite race. AU7 differences are not only due to skin color.

Next, we wanted to determine whether, next to the skin color, specific racial features of the face were responsible for the higher values for African than European faces. Therefore, we changed the light African face incrementally to a European-looking face and recorded the CV model's outputs. According to our hypothesis about which features might be more or less important for the AU7, we selected the following order: color, facial shape, eyes, forehead and brows, nose, cheeks, temples, chin, jaw, mouth. The results can be seen in Figure 10. Changing the overall shape of the face led to an important increase in detected AU activation. Changing further parts of the upper half of the face led to further but smaller increases. Changing the lower parts of the face around the mouth region did not improve the AU detection. (The small decrease is probably due to the fact that when changing one part of the face in FaceGen Modeller, the rest is not held constant but also changes slightly. Thus, previously helpful changes might change back the more regions are changed.) This means that changing the regions around the eyes to look more European lead to an increase in the detection of the lid tightening AU, while changing the regions around the mouth was not beneficial. In sum, there was not one specific feature that was responsible for the differences between races but a whole region.

Influences of perceptual characteristics and exposure to training data on the racial bias in pain detection

The next very interesting question is whether the CV model makes these differences due to a biased exposure to White European-Americans in the training stimuli or whether the differences can be attributed to purely perceptual characteristics, that is specific contrasts make specific AUs more or less easily detectable. For example, the European eye region might show some characteristics that make the detection of eye tightening easier. The fact that the CV model

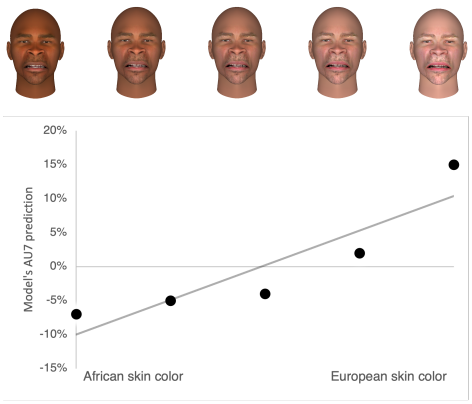


Figure 9: The CV model's AU7 prediction for the African face increases when changing the skin color.

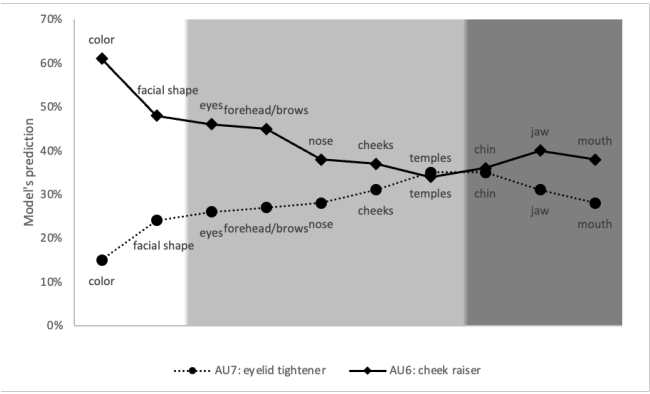


Figure 10: The CV model's prediction to African face 14 with incremental changes to make various features look more European. The white region represents changes to the whole face, the light grey region represents changes to the upper part of the face, and the darker grey region to the lower part of the face.

detected some AUs more easily in African faces and others in European faces hints that it is not the previous exposure to European-American faces that makes the detection of AUs in general easier for European faces.

Furthermore, we observed another interesting fact. AU6, the cheek raiser, and AU7, the lid tightener, share the movement of the lower lid rising upwards. Overall, the CV model's detection of AU6 was higher for African, while that of AU7 was higher in European faces. To test whether the detection of AU6 is counteracting the detection of AU7, we looked at the AU6 values of the African face 14 turning into a European face (see Figure 10). And indeed, while AU7 detection profited from making the upper part of the face look more European, AU6 detection seemed to be worsened by this. Therefore, it might be the case that the CV model detects the result of those AUs similarly well, but it attributes it differently to the specific AUs. In other words, some features in the face lead the model to interpret the facial movement either as AU6 or AU7, but it is not the case that the CV model can detect the tightened eyes per se better in one race or the other.

Conclusion

To sum up, we investigated the racial bias in pain detection with a machine learning model and artificially generated painful expressions in avatar faces. The first contribution of this work is that such generated faces are a valid method to test machine learning models. Since, most research so far has focused on real-world stimuli which contain lots of confounding variables, these datasets should be expanded by avatar faces, for which the AU activations can be controlled (and do not have to be rated by experts with varying opinions), as well as race, gender, and specific facial features. With this, one could also control the percentage of faces of different races in the training data easily. Second, and more importantly, our results show that there exist differences in the ability of our CV model to detect AUs in European vs. African faces. For some AUs, we found that this was due to the skin color difference, whereas for others the difficulty in detecting their activation was due to the characteristics of special race-dependent features in the face. This means that skin color and the characteristics of specific facial features (for example, the upper part of the face for AU7) allow for better detectability of some AUs in one race than the other. But the CV model was not consistently better to detect the activation over all AUs in one race. In some cases, like the tightness of the eyes, the outcome of specific AU activations can be detected equally well in both races, just the attribution to the specific AUs is different. All in all, our CV model does not show a strong racial bias in pain-related AU detection, even though it was mostly trained on one race. Human racial bias in pain detection can therefore not solely be explained by the visual detectability of some facial muscle activations, nor by the previous more frequent exposure to own-race faces. Other factors that are affecting humans but not pain recognition models should be investigated more closely in future research, like racially biased beliefs about pain experience (Kissi et al., 2022). Another possible explanation could be that persons with different racial and cultural backgrounds facially express pain differently. This could be investigated by using real-world pictures of painful expressions instead of artificially created ones as inputs. Furthermore, it is important to examine specific AUs in further detail, which are less easily detectable in African versus European avatar faces in order to train caretakers on which parts of the face to focus to overcome their racial bias in pain detection. To conclude, this CV approach to pain recognition and empathy is novel and enriches earlier ones that were mostly focused on its behavioral (Fabi, Weber, & Leuthold, 2019; Mende-Siedlecki et al., 2019) and neural correlates (Fabi & Leuthold, 2017; Singer et al., 2004). Future research could combine these approaches and compare the model and human performance.

Acknowledgments

Funding gratefully acknowledged from UC San Diego Social Sciences (Advancing Racial Justice award), and the Sanford

Institute for Empathy and Compassion (Center for Empathy and Technology award). In part, this project was made possible by the Science faculty of the University of Tübingen, which granted Sarah Fabi a travel scholarship. We also thank the team of FaceGen Modeller for providing us with their software and Thilo Hagendorff for helpful comments on the manuscript.

References

Andresen, N., Wöllhaf, M., Hohlbaum, K., Lewejohann, L., Hellwich, O., Thöne-Reineke, C., & Belik, V. (2020). Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *PLOS ONE*, *15*(4), 1–23.

Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020a). Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Systems with Applications*, *149*, 1–10.

Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020b). Ensemble neural network approach detecting pain intensity from facial expressions. *Artificial Intelligence in Medicine*, *109*, 1–12.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732.

Cohen, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the facial action coding system. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment.* Oxford: Oxford University Press.

De la Torre, F., Simon, T., Ambadar, Z., & Cohn, J. F. (2011). Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 57–66).

Fabi, S., & Hagendorff, T. (2022). Why we need biased AI–how including cognitive and ethical machine biases can enhance AI systems. *arXiv preprint arXiv:2203.09911*.

Fabi, S., & Leuthold, H. (2017). Empathy for pain influences perceptual and motor processing: Evidence from response force, ERPs, and EEG oscillations. *Social Neuroscience*, *12*(6), 701–716.

Fabi, S., & Leuthold, H. (2018). Racial bias in empathy: Do we process dark-and fair-colored hands in pain differently? an EEG study. *Neuropsychologia*, *114*, 143–157.

Fabi, S., Weber, L. A., & Leuthold, H. (2019). Empathic concern and personal distress depend on situational but not dispositional factors. *PLOS ONE*, *14*(11), e0225102.

Guo, Y., Wang, L., Xiao, Y., & Lin, Y. (2021). A personalized spatial-temporal cold pain intensity estimation model based on facial expression. *IEEE Journal of Translational Engineering in Health and Medicine*, *9*, 1-8.

Hjortsjö, C.-H. (1970). *Man's face and mimic language*. Lund: Studentlitteratur.

Kaltwang, S., Rudovic, O., & Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. In D. Hutchison et al. (Eds.), *Advances in Visual Computing* (pp. 368–377). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kissi, A., Van Ryckeghem, D. M., Mende-Siedlecki, P., Hirsh, A., & Vervoort, T. (2022). Racial disparities in observers' attention to and estimations of others' pain. *Pain*, *163*(4), 745–752.

Lencioni, G. C., de Sousa, R. V., de Souza Sardinha, E. J., Corrêa, R. R., & Zanella, A. J. (2021). Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. *PLOS ONE*, *16*(10), 1–12.

Lopez-Martinez, D., & Picard, R. (2018). Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In *International Conference of the IEEE Engineering in Medicine and Biology Society* (p. 5624-5627).

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. In *IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 57–64).

Martinez, D. L., Rudovic, O., & Picard, R. (2017). Personalized automatic estimation of self-reported pain intensity from facial expressions. In Y. Liu, J. M. Rehg, C. J. Taylor, & Y. Wu (Eds.), *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 2318–2327). IEEE.

Mays, V. M., Cochran, S. D., & Barnes, N. W. (2007). Race, race-based discrimination, and health outcomes among African Americans. *Annual Reviews of Psychology*, *58*, 201–225.

Mende-Siedlecki, P., Qu-Lee, J., Backer, R., & Van Bavel, J. J. (2019). Perceptual contributions to racial bias in pain recognition. *Journal of Experimental Psychology*, *148*(5), 863.

Noor, A., Zhao, Y., Koubaa, A., Wu, L., Khan, R., & Abdalla, F. Y. (2020). Automated sheep facial expression classification using deep transfer learning. *Computers and Electronics in Agriculture*, *175*, 1–8.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Pouromran, F., Radhakrishnan, S., & Kamarthi, S. (2021). Exploration of physiological sensors, features, and machine learning models for pain intensity estimation. *PLOS ONE*, *16*(7), 1–17.

Prkachin, K. M., & Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, *139*(2), 267–274.

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology*, 3–24.

Sessa, P., Meconi, F., Castelli, L., & Dell'Acqua, R. (2014). Taking one's time in feeling other-race pain: An event-related potential investigation on the time-course of cross-

racial empathy. *Social Cognitive and Affective Neuroscience*, *9*(4), 454–463.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162.

Soar, J., Bargshady, G., Zhou, X., & Whittaker, F. (2018). Deep learning model for detection of pain intensity from facial expression. In M. Mokhtari, B. Abdulrazak, & H. Aloulou (Eds.), *Smart homes and health telematics, designing a better future: Urban assisted living* (pp. 249–254). Cham: Springer International Publishing.

Susam, B. T., Akcakaya, M., Nezamfar, H., Diaz, D., de Sa, V. R., Craig, K. D., ... Goodwin, M. S. (2018). Automated pain assessment using electrodermal activity data and machine learning. In *International conference of the IEEE Engineering in Medicine and Biology Society* (p. 372-375).

Susam, B. T., Riek, N. T., Akcakaya, M., Xu, X., de Sa, V. R., Nezamfar, H., ... Huang, J. S. (2021). Automated pain assessment in children using electrodermal activity and video data fusion via machine learning. *IEEE Transactions on Biomedical Engineering*, *69*(1), 422-431.

Tavakolian, M., Bordallo Lopez, M., & Liu, L. (2020). Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recognition Letters*, *140*, 26–33.

Tavakolian, M., & Hadid, A. (2018). Deep binary representation of facial expressions: A novel framework for automatic pain intensity recognition. In *IEEE International Conference on Image Processing* (p. 1952-1956).

Tavakolian, M., & Hadid, A. (2019). A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics. *International Journal of Computer Vision*, *127*(10), 1413–1425.

Thuseethan, S., Rajasegarar, S., & Yearwood, J. (2019). Deep hybrid spatiotemporal networks for continuous pain intensity estimation. In T. Gedeon, K. W. Wong, & M. Lee (Eds.), *Neural Information Processing* (pp. 449–461). Springer International Publishing.

Xin, X., Lin, X., Yang, S., & Zheng, X. (2020). Pain intensity estimation based on a spatial transformation and attention CNN. *PLOS ONE*, *15*(8), 1–15.

Xu, X., Craig, K., Diaz, D., Goodwin, M., Akcakaya, M., Susam, B., ... de Sa, V. R. (2018). Automated pain detection in facial videos of children using human-assisted transfer learning. In *Artificial Intelligence in Health - 2019: First International Workshop, AIH 2018. Revised Selected Papers. LNAI 11326* (p. 162-180).

Xu, X., & de Sa, V. R. (2020). Exploring Multidimensional Measurements for Pain Evaluation using Facial Action Units. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (p. 559-565).

Xu, X., & de Sa, V. R. (2021). Personalized pain detection in facial video with uncertainty estimation. In *International Conference of the IEEE Engineering in Medicine & Biology Society* (pp. 4163–4168).

Xu, X., Huang, J. S., & de Sa, V. R. (2020). Pain evaluation in video using extended multitask learning from multidimensional measurements. In *NeurIPS: Machine Learning for Health Workshop* (pp. 141–154).

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience*, *29*(26), 8525–8529.

Zafar, Z., & Khan, N. A. (2014). Pain intensity evaluation through facial action units. In *2014 22nd international conference on pattern recognition* (p. 4696-4701).

Zhou, J., Hong, X., Su, F., & Zhao, G. (2016). Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*