Journal of Behavioral Data Science, 2025, *5* (2), 1–15.
DOI: https://doi.org/10.35566/jbds/wyman51

Evaluating the Threat of Phantom Faces in Emotion Detection AI through Simulation

Austin Wyman[1] and Zhiyong Zhang[1]

[1]Department of Psychology, University of Notre Dame, Notre Dame, USA
awyman@nd.edu

Abstract. Emotion detection AI is an emerging tool in the field of psychology that enables researchers to process large batches of images of human faces and obtain estimates of the emotions present within images. Some algorithms, such as Py-Feat, are even capable of detecting multiple faces within an image and providing differential estimates for each face. However, a known problem with multiple detection algorithms is that they sometimes mistakenly detect multiple faces when only a single face exists. In such cases, detection of the true face is still available to users and the false face can be ignored, but there may be artifacts of the false face within the true face that are biasing the estimation of emotions. The present study investigated whether the presence of a second face reduces the accuracy of emotion estimation in the first face. Using 1,438 images from the RAVDESS labeled emotion data set, we generated image with multiple faces under a variety of conditions (i.e., size, opacity, emotion similarity, and number of faces) and compared them against unaltered, single face versions of the images. There were meaningful differences in accuracy across between the single-face and multiple-face images, with similarity and number of faces being the most detrimental conditions for multiple-face accuracy. Findings suggest that it is highly important for researchers to remove extraneous faces within images in order to maximize the accuracy of emotion detection analysis.

Keywords: Emotion Detection · Emotion Recognition · Artificial Intelligence · Distortion · Phantom Faces · Multiple Faces

1 Introduction

1.1 Introduction to emotion detection AI

Emotion detection AI (also known as emotion recognition API) is an emerging application of artificial intelligence (AI) used to detect, label, and understand human emotions from images and videos. The technology has been applied in a variety of clinical and educational research settings (Wyman & Zhang, 2023).

Within clinical research, emotion detection AI is often used to develop automated interventions, which monitor participants' emotions and respond with stimuli to influence behavior (Alharbi & Huang, 2020; Bharatharaj, Huang, Mohan, Al-Jumaily, & Krägeloh, 2017; Grossard et al., 2017; Jiang et al., 2019; Liu, Wu, Zhao, & Luo, 2017; Manfredonia et al., 2018). Within educational research, the technology has been used to monitor emotions in response to educational interventions, such as online learning (Chu, Tsai, Liao, & Chen, 2017; Chu, Tsai, Liao, Chen, & Chen, 2020), which are concurrently taking place. However, emotion detection AI is not limited to these applications. In fact, several disciplines in the social and behavioral sciences could benefit from its implementation. Emotion detection AI itself is the integration of research across multiple disciplines, including psychology, physiology, and computer science (Wyman & Zhang, 2025). The technology is based on the concept of action units (AUs, Ekman & Friesen, 1976), which are the simplest combinations of muscles required to produce a facial expression. For example, AU4 corresponds to the act of lowering one's brow and requires the depressor glabellae, depressor supercilli, and corrugator supercilli—three muscles located in the forehead. The Facial Action Coding System (FACS, Ekman & Friesen, 1978) assigns basic emotions to the combination of AUs. For example, when AU4 "brow lowerer" is combined with "upper lid raiser" (AU5), "lid tightener" (AU7), and "lip tightener" (AU23), the facial expression for anger is produced. The FACS traditionally included six emotions—happiness, sadness, anger, surprise, disgust, and fear—but future models were extended to include more emotions like contempt and confusion.

Modern emotion detection AI operationalizes the FACS through a two-step convolutional neural network (CNN), in which the first step of the network focuses on face recognition and the second step on emotion classification. CNNs are a class of artificial neural networks that specialize in processing grid-like topology (Baduge et al., 2022), such as image data, which are treated as a two-dimensional grid of pixels. CNNs are often used to identify patterns within image, such as to detects edges in shapes (Dorafshan, Thomas, & Maguire, 2018), transcribe text from images (Wei, Sheikh, & Ab Rahman, 2018), or recognize faces (Lawrence, Giles, Tsoi, & Back, 1997). CNNs are uniquely suited for processing image data because of their aptitude for handling sparsity. Neural networks are powerful prediction models because they are able to handle multiple layers of parameters that explain complex, often non-linear relationships in the data. However, estimating thousands to millions of parameters when only tens to hundreds are meaningful is computationally expensive (Goodfellow, Bengio, Courville, & Bengio, 2016). The problem is particularly defined for image data, as traditional neural networks are inefficient to the handle the sparse data caused by background and non-focal pixels. CNNs address this problem through a convolution step, which obtains summaries of pixels given by their surrounding information and prioritizes pools of pixels with the most information. In Py-Feat and similar emotion detection AI models (Wyman & Zhang, 2025), the purpose of CNNs is to identify the location of facial features in an image, which is fed to a secondary neural network to determine if action units are activated. Finally, a probabilis-

tic model is conducted, which estimates the probability that a given emotion is being observed given the activated action units. Some emotion detection AI models provide a discrete classification based on the emotion with the highest probability estimate, assuming that an image can only depict one emotion at a time. Other models assume that humans exhibit multiple emotions at once, providing the raw probability estimates for each emotion. Although, different emotion detection AI models use the output in different ways, they all adhere to the same CNN architecture.

1.2 Introduction to Py-Feat

Another difference between emotion detection AI models is whether the model is open-source or commercially-based, which impacts the amount of pre-trained data that is available and the degree of user customizability (Wyman & Zhang, 2025). The Python Facial Expression Analysis Toolbox (Py-Feat, Cheong et al., 2023) is emerging as a valuable open-source model for emotion detection AI, which was created by psychologists for psychologists. The toolbox features 7 emotions (happiness, sadness, anger, surprise, disgust, fear, and neutral) and each emotion is rated continuously on a 0-1 decimal scale. It allows for multiple emotions per image, with each emotion rating representing the proportion of the total face that is exhibiting the given emotion. The Py-Feat architecture consists of five building blocks, which represent different steps of the facial expression analysis procedure. Each block is controlled by a pre-trained, open-source model and can be exchanged by the user for a different model. By default, Py-Feat provides one pre-trained model for face and facial pose estimation, three for facial landmark detection, two for action unit detection, two for emotion detection, and one for identity detection. In particular, the default emotion detection model is the Residual Masking Network (ReMaskNet, Pham, Vu, & Tran, 2021), which Cheong et al. (2023) demonstrated performs better on images in the wild than some commercial emotion detection AI models like iMotions.

Most emotion detection AI models are evaluated using posed images, or images that whose lighting, positioning, and background are carefully designed as to not disrupt the algorithm. However, posed images are not realistic representations of how emotion detection AI models are used in the field, which is why most models have reduced accuracy for images in the wild. Py-Feat has been validated under benchmarked datasets of images in the wild and also has completed robustness tests against common image barriers to facial expression analysis, such as luminance, occlusion, and head rotation (Cheong et al., 2023). Luminance describes the impact of various lighting conditions, either extreme brightness or darkness, which may inhibit the detection of facial landmarks and AUs. Cheong et al. (2023) found that Py-Feat was robust against issues related to luminance on both ends of the spectrum. Occlusion describes the partial obstruction of facial features by an object blocking the face, which similarly inhibits the detection of facial landmarks and AUs. Py-Feat encounters a substantial decline in performance in response to face occlusion (Cheong et al., 2023). Accuracy for face detection, AU detection, and emotion detection models all declined if either

the eyes, nose, or mouth of an image were hidden. Finally, head rotation refers to the direction in the which face is facing the camera in an image. Models are often trained with faces that directly face the camera, but images in the wild are rarely facing straight forward. Models often have challenges detecting the facial landmarks and AUs of side-facing images, which may not generalize to training data. Py-Feat demonstrated robustness against the issue of head rotation (Cheong et al., 2023). The toolbox is a valuable resource for facial expression analysis, as it has been trained and validated on images in the wild, which are more representative of actual usage.

1.3 Phantom faces in emotion detection AI

Aside from traditional image distortions (i.e., luminance, occlusion, and rotation), there is a rare image distortion that has been observed by users of Py-Feat but has not been formally documented in the literature. The distortion is related to the face and facial pose estimation component of Py-Feat and it incorrectly identifies a secondary face within the true face of an image, often located on the forehead of the true face. Figure 1 presents an example of the issue with an image from a benchmarked dataset (Livingstone & Russo, 2018). Note that Py-Feat produces confidence estimates for each face that it detects. The primary face is detected with a confidence of 99.9%, whereas the second, false face is detected with a confidence of 79.9%. The high confidence to detect a second face is concerning given the lack of a second face altogether. No literature exists to define the issue of two faces, nor does it offer any explanation. Hence, given its apparitional appearance, we refer to the issue as "phantom faces".

An easy solution to phantom faces is adjusting the threshold for face detection. For example, the phantom face in the example image had a confidence of 79.9%. By setting the confidence threshold to 0.8 or higher, no analysis would be conducted for any faces below the confidence threshold. However, this solution ignores the issue rather than solving it, as artifacts of the phantom face may remain within the true face even after filtering it out. It is difficult to remove the influence of the phantom face from an image without knowledge of what causes the issue. Therefore, the priority of research should be to identify the extent to which phantom faces bias estimation of emotions in the true faces. Moreover, the issue of biased estimation extends to other cases of emotion detection with multiple faces. Given the frequency of multiple faces in real-world images, often as figures in the background of landscapes or experiments, it is important to understand the extent to which non-focal faces bias the primary face. An empirical examination of the impact of phantom faces and multiple faces on emotion estimation would greatly improve the experimental considerations and practices regarding emotion detection AI and improve the quality of research published in the field.

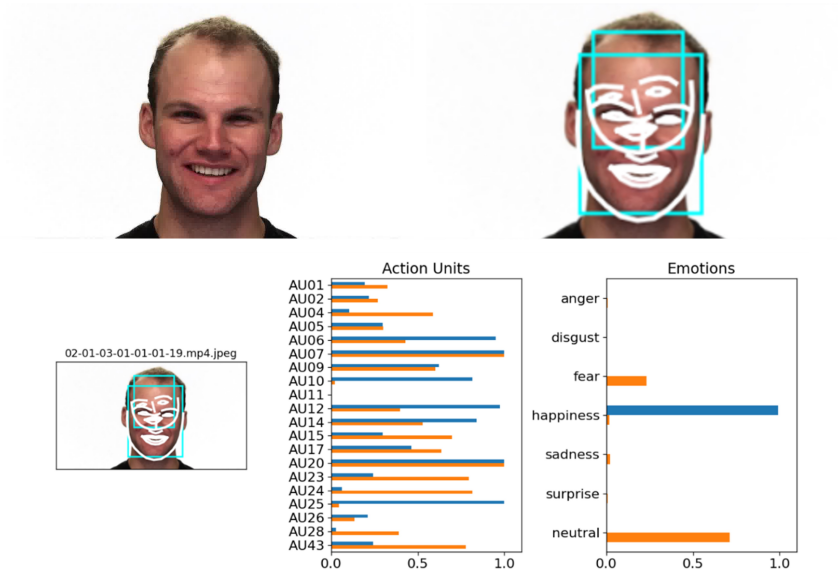

Figure 1. Example of "phantom face" distortion appearing within an image and its corresponding emotion detection AI output. Note. (top left) Original image (02-01-03-01-01-01-19) sampled from a video in the RAVDESS dataset. (top right) Image after processing by Py-Feat, identifying two faces. (bottom) Complete Py-Feat output with blue bars indicating the first identified face (true face) and orange indicating the second face (phantom face). The true face is correctly identified as happy whereas the phantom face is identified as neutral, with low probability of fear, happiness, and sadness.

1.4 Model evaluation for emotion detection AI

Given that the issue of phantom faces has not been discussed in the literature, there is no existing framework for evaluating emotion detection AI models with respect to phantom faces. Currently, models are evaluated using labeled image datasets, which specify a correct response that emotion detection models should be able to match. For example, the dataset may include an image labeled "happy" and for the model is get the case correct it must also produce a "happy" label. Models are evaluated by their accuracy, or how many labels they can correctly match, and their accuracy under various conditions. The conditions are often artificially induced by editing the labeled image. For example, Cheong et al. (2023) created artificial occlusion in images by editing a black bar to cover either the eyes, ears, or mouth of the subject in the image, and manipulated the brightness of the image with a filter to simulate luminance conditions. Yang et al. (2021) similarly applied a Gaussian Blur to images to simulate motion blur and noise from cameras. Some studies also evaluate emotion detection AI models using benchmarked datasets that are designed to include images containing distortions (Kuruvayil & Palaniswamy, 2022; Mollahosseini, Hasani, & Mahoor,

2017). The approach is simple and can be easily replicated across multiple studies. However, using a benchmarked dataset is only accessible when there exists a large collection of images with the intended distortion. When images do not exist, it is necessary to design an experiment and recruit participants, which can be expensive. Moreover, the process of labeling new image data can be onerous based on the large sample size necessary to make stable inferences regarding model performance. Simulating distortions in images is more accessible to answer certain questions related to emotion detection AI model evaluation.

1.5 Present study

Currently, there is no existing benchmarked dataset that describes phantom faces or any issues related to multiple face detection, meaning the question of their impact on emotion detection AI models must be evaluated through a simulated data. The purpose of the present study is to understand the risk of phantom faces or multiple faces in classification tasks using emotion detection AI. Distortions like facial occlusion reduce the accuracy of emotion detection AI models by blocking the estimation of AUs. Phantom faces may also block necessary facial landmarks and interfere with AUs. Therefore, the primary hypothesis is that the presence of phantom faces leads to a decrease in accuracy in emotion classification. To study this, the present study develops a novel experiment for simulating phantom faces, which may be replicated by other researchers evaluating emotion detection AI models. Emotion detection AI may be a significant technology for advancing the emotion research in the social and behavioral sciences, but its success is dependent on the support of rigorous frameworks for model evaluation.

2 Methods

2.1 Materials and procedures

The present study utilized image data from the publicly available Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS, Livingstone & Russo, 2018) dataset. RAVDESS contains 7,356 video and audio recordings. The dataset contains both speech and song recordings, but the present study only utilizes the subset of speech recordings. There are 4,320 speech recordings available, which are divided into full audio-video, video only, and audio only recordings. Only the 1,440 video only recordings were relevant to emotion detection. During the experiment, actors were instructed to vocalize two statements ("Kids are talking by the door", "Dogs are sitting by the door") with eight emotional intentions (neutral, calm, happy, sad, angry, fearful, surprise, and disgust) and two emotional intensities (normal, strong); however, conditions involving the neutral emotion were only vocalized with normal intensity. Actors repeated each vocalization twice, resulting in 60 total trials per actor ($N = 24$).

Actors were recruited from the Toronto, Canada area with ages ranging from 21 to 33. Odd-numbered actors were male ($n = 12$) and even-numbered actors

were female ($n = 12$). Additionally, actors represented Caucasian ($n = 20$), East-Asian ($n = 2$), and multiracial ($n = 2$) ethnic backgrounds. Actors did not have any distinctive facial features, such as facial hair or tattoos, and were instructed not to wear glasses, in order to ensure minimal interference with face detection algorithms. There are multiple benchmarking datasets available for emotion detection AI, but the RAVDESS dataset was selected as an appropriate dataset because of its highly structured system of labeled data, which allow researchers to evaluate the accuracy of models at various conditions (e.g., statement, intensity).

The RAVDESS dataset only includes videos of participants. Images were extracted as frames from each trial using the "av" package in R (Ooms, 2024a), sampling at a frame rate of three frames per second. The sampling produced an average of 11 images per video with a standard deviation of one image. Although all images in a video share the same label, the images may not be equally representative of the label. For example, a video may contain the emotion anger but the actor may not be presenting anger the entire time, as there may be resting emotions also captured on camera before or after. Therefore, it is important to isolate the most representative frame of the target emotion (i.e., the emotion labeled by the dataset). Py-Feat emotion estimates were obtained for each image within a trial. The image with the maximum estimate for the target emotion was selected as the most representative frame of the video and used for the present analysis. The final sample contained 1,246 images, which consisted of the most representative frames from the 1,440 videos excluding the trials labeled as "calm", given that Py-Feat is not able to detect calm emotions.

2.2 Simulation design

The present study created control and experiment sets of images based on the RAVDESS sample. The set of control images were unedited from the original RAVDESS images, depicting a singular face. The set of experiment images were identical to the control set except that they were edited to include an additional face. Within the experiment set, the control image was the primary focus of the camera and an additional face was selected from the RAVDESS sample to appear somewhere in the background close to the face, but without occlusion of the face. Phantom faces were inserted into the experiment set using the "magick" package in R (Ooms, 2024b) and were edited according to a variety of conditions: size, opacity, number, and sameness. The size condition describes the size of the phantom face in the image, with the phantom face appearing as either 25% or 50% of the size of the focal image. The opacity condition describes the lack of transparency of the phantom face, with the phantom face appearing fully visible at 0% opacity, half visible at 50% opacity, and near completely transparent at 75% opacity. The number condition describes the number of phantom faces that appears in the image. The number condition 1 inserts a first phantom face into the top left corner of the image, 2 inserts a second phantom face in the top right, 3 inserts the third face in the bottom left, and 4 inserts the fourth in the bottom right. A location condition is somewhat nested within the number condition, but

the location of phantom face should not matter as long as the phantom face is equally obstructing the primary face in all four conditions. Phantom faces are carefully positioned as to avoid occlusion; therefore, location is not a meaningful condition. Finally, the sameness condition determines whether the phantom face exhibits the same emotion as the primary face and, if the emotions are different, which emotion is exhibited. The values of the sameness condition are "same", "anger", "disgust", "fear", "happiness", "neutral", "sadness", and "surprise." In the "same" condition, phantom faces inserted were identical to the primary face. In the remaining emotion conditions, phantom faces that were most representative of the target emotion were inserted. The most representative images were again identified from RAVDESS and were selected by which image the maximum target emotion estimates in the sample. The present study explored 192 total simulation conditions.

2.3 Data analysis

The present study examined the difference in performance of Py-Feat on images in the control set and experiment set. Performance was primarily measured by the overall classification accuracy of each image set, parameter bias in emotion estimation, and the conditional classification accuracy within each simulation condition. Differences in overall accuracy were evaluated using a paired t-test. Given the similarity of images in the control and experiment set, a paired t-test is appropriate because the difference in accuracy is approximately normal at large samples and sample variances are approximately equal. A similar t-test approach was used to evaluate parameter bias in emotion estimates for each emotion label across the 192 conditions. Finally, differences in conditional accuracy were evaluated using an Analysis of Variance (ANOVA) model to investigate the main effects of each variable and their two-way interaction effects. All analyses were conducted in R version 4.4.1.

3 Results

Py-Feat correctly identified labeled emotions in 81.3 percent of control set images. Since the control set was not subjected to any simulation conditions, the control set accuracy was consistent across all conditions. Py-Feat demonstrated substantially lower accuracy in the experiment set of images. The average accuracy for the experiment set was 54.7 percent with a standard deviation of 18.5 percent. The average difference in accuracy between the control and experiment set was 26.6 percent, which the paired t-test demonstrated was statistically significant, $t(191) = 19.9, p < .001$. Moreover, Cohen's d was 1.44, indicating that the presence of phantom faces in images substantially impacts Py-Feat's overall classification accuracy.

Py-Feat's classification accuracy is also impacted by the condition of phantom face images. The main effect of size on the difference in accuracy between control and experiment sets of images was statistically significant, $F(1, 190) =$

89.8, $p < .001$. Py-Feat performed worse at labeling emotions with phantom faces that were 50% of the size of the primary face (Mean difference = 0.37, SD = 0.18) than phantom faces that were 25% (M = 0.16, SD = 0.12). The main effect of sameness on accuracy difference was also significant, $F(7, 184) = 37.5, p < .001$. When phantom images exhibited the same emotion as the primary face, there was a small difference in accuracy between the control and experiment set (M = 0.05, SD = 0.06). However, there were more pronounced differences when the primary and phantom faces exhibited different images. The largest difference was observed when the phantom face exhibited happiness (M = 0.49, SD = 0.15), followed by anger (M = 0.42, SD = 0.15), fear (M = 0.37, SD = 0.13), disgust (M = 0.25, SD = 0.09), neutral (M = 0.23, SD = 0.04), surprise (M = 0.18, SD = 0.16), and sadness (M = 0.13, SD = 0.12). The interaction effect of size and sameness was also statistically significant, $F(7, 176) = 41.6, p < .001$, indicating that the magnitude of bias caused by the size of phantom images varied depending on the emotion of the phantom face. Table 1 presents a summary of the Tukey Honest Significance Difference test comparisons, which examined the simple effects of the interaction. The difference between the 25% and 50% conditions was greatest within the happiness and surprise conditions, but the 50% condition performed significantly worse across the different levels of the sameness condition. In contrast, the main effects of opacity, $F(2, 189) = 0.5, p = 0.542$, and number, $F(3, 188) = 0.1, p = 0.985$, were not statistically significant and nor were their interaction effects with any of the other variables. Table 2 presents the highest and lowest differences in accuracy conditions, corroborating the claim that phantom face accuracy is largely determined by size and sameness.

Table 1. Tukey Honest Significance Difference test contrasts among size conditions within sameness conditions.

Emotion	Mean difference (SE)	t statistic
Same	-0.076 (0.014)	-5.25 ***
Anger	-0.274 (0.014)	-18.79 ***
Disgust	-0.172 (0.014)	-11.81 ***
Fear	-0.256 (0.014)	-17.58 ***
Happiness	-0.303 (0.014)	-20.78 ***
Neutral	-0.071 (0.014)	-4.89 ***
Sad	-0.22 (0.014)	-15.06 ***
Surprised	-0.30 (0.014)	-20.71 ***

Note. *** $p < .001$. All t statistics were obtained with 176 degrees of freedom.

Differences in continuous emotion estimates between images in the control and experiment set were also examined to examine the effect of phantom faces on parameter estimation. Each emotion was examined separately. All means are expressed in units of Py-Feat estimates, which range from 0-1. The largest difference was observed among anger estimates, $d = -0.15$. There was a -0.04 mean difference in anger estimates, $t(276095) = -77.6, p < .001$, indicating that Py-

Table 2. Best case and worst case scenarios for phantom face conditions.

Case Number	Difference	Size	Opacity	Number	Emotion
1	0.006	25%	0%	1	Same
2	0.007	25%	0%	2	Same
3	0.007	25%	0%	4	Same
190	0.651	50%	50%	3	Happiness
191	0.651	50%	50%	2	Happiness
192	0.652	50%	0%	3	Happiness

Feat overestimates the anger of images when phantom faces are present. A 0.04 mean difference in neutral estimates was observed, $t(276095) = 43.7, p < .001$. The effect size was positive, $d = 0.14$, indicating that Py-Feat underestimates neutral emotions in the presence of phantom faces. There was a -0.03 mean difference in happiness estimates, $t(276095) = -47.2, p < .001$, with an effect size of $d = -0.09$. There was a 0.02 mean difference in surprise estimates, $t(276095) = 43.7, p < .001$, with an effect size of $d = 0.08$. There was a 0.01 mean difference in sadness estimates, $t(276095) = 33.3, p < .001$, with an effect size of $d = 0.06$. There was a -0.01 mean difference in fear estimates, $t(276095) = -23.9, p < .001$, with an effect size of $d = -0.04$. Finally, there was a 0.008 mean difference in disgust estimates, $t(276095) = 14.8, p < .001$, with an effect size of $d = 0.03$. All differences were statistically significant, but were classified as small effect sizes.

4 Discussion

4.1 Findings and implications

Py-Feat demonstrated lower accuracy at classifying images in the experiment set than the control set, and the difference was statistically significant. However, it is a known property of statistical tests that as sample size becomes increasingly large, statistical tests will always converge toward a statistically significant result, regardless of how menial the practical significance of the result is (Meehl, 1967). Therefore, we should prioritize the effect size of results because it is unaffected by sample size. The effect size of the first paired t-test was 1.44, which is substantially larger than Cohen's threshold for a large effect size (0.8). Therefore, we are confident that the observed difference in accuracy between the control and experiment set is practically significant as well. The difference in overall classification accuracy is sufficient to claim that phantom faces are a valid threat to the inference of emotion detection AI. Py-Feat users that conduct analysis on images that contain phantom faces can expect a substantial reduction in accuracy.

However, the effect of bias was not uniformly distributed across the simulation conditions. Some phantom faces conditions, such as 25% size and same emotion, resulted in less than a 1 percent difference in accuracy between the

control set and experiment set. Other conditions, such as 50% size and happiness, resulted in a massive 65 percent difference in accuracy. Opacity had no main or interaction effects with accuracy difference, suggesting that Py-Feat's algorithm is sophisticated enough to detect facial landmarks and AUs regardless of how transparent the image is. It seems that as long as Py-Feat is able to detect the phantom face at all, the phantom face biases the image. Similarly, neither the number main effect nor its interactions were significant, suggesting that there is no multiplicative impact of multiple phantom faces. The presence of one phantom face alone is enough to bias the image. The lack of a significant effect associated with number also suggests that there is no significant effect of phantom face location, as long as the phantom face is not directly obstructing the primary face. However, it is important to note that there were no safe conditions observed. The presence of a phantom face in the experiment set always produced lower accuracy than their control set baseline, yet various conditions determined the severity of the bias that was observed.

Py-Feat consistently performed better in the presence of smaller phantom faces than larger phantom faces, but the impact was differentially observed for different clusters of emotions. The smallest difference was observed for same and neutral emotions, which is an intuitive result. In the same condition, the phantom face was a duplicate of the primary face, except smaller, and therefore, it was only capable of biasing the primary face with its own AUs. The only bias that could be produced by the same condition is that which is caused by occlusion, which was intentionally limited in the experiment design. The neutral emotion is unique from other motions in the FACS because it is defined by the lack of any AUs activated at all. Therefore, AUs from a neutral phantom face were not able to interfere with the AUs of the primary face, as the phantom face AUs did not exist. The larger accuracy difference due to happiness can also be explained by AUs. Happiness is one of the simplest facial expressions to explain by AUs, consisting of only two AUs: "cheek raiser" (AU6) and "lip corner puller" (AU12). Additionally, the two AUs are not repeated in any other emotion, making the presence of the two AUs with any combination of other AUs an easy decision to label the image as happiness. Consequently, emotion detection AI models tend to have the highest accuracy classifying happiness labels, with some models even achieving 100 percent accuracy in benchmarked datasets (Yang et al., 2021). Therefore, it is likely that Py-Feat defaults to classifying happiness emotions, which it has the highest accuracy for, when it detects the necessary AUs in the phantom face, even if they do not appear in the primary face. The next large decrease in accuracy was caused by the surprise condition, which is an emotion that shares multiple AUs with fear and anger—two other emotions that resulted in a large decrease. It is likely that Py-Feat confused the three emotions because of their similarity, as it combined AUs from both the primary face and phantom face. The difference between the three emotions may have been more pronounced at larger phantom face sizes than smaller because the AUs were available in a higher resolution, making them easier to detect.

Paired t-test results for the continuous estimates were all statistically significant, but it is not recommended to rely on statistical significance given the large number of simulation cases ($N = 276095$). The effect size estimates paint a different story, finding that the true differences in emotion estimates were all approximately 0. The small effect is likely due to the aggregation across all simulation cases, including cases in which neither primary face nor the phantom face exhibited the emotion of interest. However, it is important to note that the difference in continuous estimates was not unidirectional. Anger, happiness, and fear were overestimated by Py-Feat in the presence of phantom faces, whereas neutral, surprise, sadness, and disgust were underestimated. The results corroborate the previous claims that happiness estimates are detected more often because of their AUs and that anger and fear are often mistaken for surprise in phantom faces. However, the effect of phantom faces conditions on continuous estimates remains unknown, which may influence the severity of bias.

4.2 Limitations and future directions

The present study contributed a novel experiment design framework for evaluating the issue of phantom faces in emotion detection AI; however, there are multiple limitations to the current design. Phantom faces cannot be directly replicated in an image because what causes phantom faces to appear naturally is unknown. Therefore, it is uncertain whether the experiment set of images is representative of phantom faces encountered in the wild, but it certainly is generalizable to the broader problem of multiple faces in emotion detection. Multiple faces appear in images under a variety of circumstances, whether they are a passing figure in the background or an active backdrop of an experiment (e.g., classrooms). The present study identified the risk of including any non-primary face in the background of emotion detection tasks, which results in a substantial decrease in classification accuracy and an increase in bias for continuous emotion estimation. Researchers conducting emotion detection AI work should prioritize removing the influence of any non-primary faces from the image before conducting any analysis.

The present study introduced the issue of phantom faces and examined its risks, but it did not provide any empirical solutions for addressing the problem. Researchers could crop images around the primary face to remove the influence of other faces. However, cropping images would not address phantom faces that appear within the primary face. Additionally, cropping images may not be feasible when important information is contained within the background of an image. Future research should investigate other methods for removing the influence of other faces, such as background blurring or targeted face blurring, which may not have such tradeoffs.

Another limitation of the present simulation design is that the experiment set of images was only evaluated under four variables, some of which containing only 2 or 3 levels. Future research may want to investigate other size and opacity parameters than the ones selected in the present simulation. Additionally, there may be other factors that influence the severity of phantom faces, which

were not considered in the present study. Future research should identify these factors and expand the simulation paradigm of emotion detection AI to evaluate its robustness across a diversity of conditions. Finally, this study only evaluated the performance of Py-Feat but several other emotion detection AI models are available, such as Amazon Rekognition and Google Cloud AI, and can be investigated in the future. Future research should observe how the problem of phantom faces replicates across other models and the novel solutions that may emerge.

4.3 Conclusion

Emotion detection AI is an emerging technology in the social and behavioral sciences, which may transform the accessibility of multimodal designs in emotion research; however, the current technology is limited by the lack of rigorous methodology for AI model evaluation. Simulation studies using edited images revealed insight into the problem of phantom faces and multiple faces, but they may provide insight into other challenges with emotion detection AI models as well. The paradigm of simulation studies has bolstered quantitative methodology by elucidating the circumstances in which methods flounder or flourish. It can be applied just as eagerly to AI model evaluation, provided that at ground truth is known, such as with labeled data sets. The present simulation introduced the presence of phantom faces as a substantive issue, which we hope motivates other researchers to identify possible solutions. Through the continuation of rigorous evaluation work, emotion detection AI may become a valuable tool for emotion research.

Acknowledgments

The study was presented at the 2025 Annual Meeting of the International Society for Data Science and Analytics. Wyman is supported by the NSF Graduate Research Fellowship (2236418), the Notre Dame Program for Interdisciplinary Education Research Burns Fellowship, and the Lucy Graduate Scholars Program. Zhang is supported by the US Department of Education (R305D210023) and Notre Dame Global. The authors certify that they have no conflicts of interests to declare that are relevant to the content of this article.

References

Alharbi, M., & Huang, S. (2020). An augmentative system with facial and emotion recognition for improving social skills of children with autism spectrum disorders. In *2020 IEEE International Systems Conference (SysCon)* (pp. 1–6). doi: https://doi.org/10.1109/SysCon47679.2020.9275659

Baduge, S. K., Thilakarathna, S., Perera, J. S., Arashpour, M., Sharafi, P., Teodosio, B., ... Mendis, P. (2022). Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. *Automation in Construction*, *141*, 104440. doi: https://doi.org/10.1016/j.autcon.2022.104440

Bharatharaj, J., Huang, L., Mohan, R. E., Al-Jumaily, A., & Krägeloh, C. (2017). Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder. *Robotics*, *6*(1), 4. doi: https://doi.org/10.3390/robotics6010004

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., & J, C. L. (2023). Py-Feat: Python facial expression analysis toolbox. *Affective Science*, *4*, 781–796. doi: https://doi.org/10.1007/s42761-023-00191-4

Chu, H.-C., Tsai, W.-H., Liao, M.-J., & Chen, Y.-M. (2017). Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning. *Soft Computing*, *22*, 2973–2999. doi: https://doi.org/10.1007/s00500-017-2549-z

Chu, H.-C., Tsai, W.-H., Liao, M.-J., Chen, Y.-M., & Chen, J.-Y. (2020). Supporting e-learning with emotion regulation for students with autism spectrum disorder. *Educational Technology & Society*, *23*(4), 124–146. Retrieved from https://www.jstor.org/stable/26981748

Dorafshan, S., Thomas, R. J., & Maguire, M. (2018). Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Construction and Building Materials*, *186*, 1031–1045. doi: https://doi.org/10.1016/j.conbuildmat.2018.08.011

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. Environmental psychology & nonverbal behavior. *Journal of Personality and Social Psychology*, *1*(1), 56–75. doi: https://doi.org/10.1007/BF01115465

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system.* American Psychological Association (APA). doi: https://doi.org/10.1037/t27734-000

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning.* Cambridge: MIT Press.

Grossard, C., Grynspan, O., Serret, S., Jouen, A.-L., Bailly, K., & Cohen, D. (2017). Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (ASD). *Computers & Education*, *113*, 195–211. doi: https://doi.org/10.1016/j.compedu.2017.05.002

Jiang, M., Francis, S. M., Srishyla, D., Conelea, C., Zhao, Q., & Jacob, S. (2019). Classifying individuals with asd through facial emotion recognition and eye-tracking. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6063–6068). doi: https://doi.org/10.1109/EMBC.2019.8857005

Kuruvayil, S., & Palaniswamy, S. (2022). Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. *Journal of King Saud University-Computer and Information Sciences*, *34*(9), 7271–7282. doi: https://doi.org/10.1016/j.jksuci.2021.06.012

Lawrence, S., Giles, C., Tsoi, A. C., & Back, A. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113. doi: https://doi.org/10.1109/72.554195

Liu, X., Wu, Q. J., Zhao, W., & Luo, X. (2017). Technology-facilitated diagnosis and treatment of individuals with autism spectrum disor-

der: An engineering perspective. *Applied Sciences*, *7*(10), 1051. doi: https://doi.org/10.3390/app7101051

Livingstone, S. R., & Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, *13*(5), e0196391. doi: https://doi.org/10.1371/journal.pone.0196391

Manfredonia, J., Bangerter, A., Manyakov, N. V., Ness, S., Lewin, D., Skalkin, A., ... others (2018). Automatic recognition of posed facial expression of emotion in individuals with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *49*, 279–293. doi: https://doi.org/10.1007/s10803-018-3757-9

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of science*, *34*(2), 103–115. doi: https://doi.org/10.1086/288135

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. doi: https://doi.org/10.1109/TAFFC.2017.2740923

Ooms, J. (2024a). av: Working with audio and video in r [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=av` (R package version 0.9.3)

Ooms, J. (2024b). magick: Advanced graphics and image-processing in r [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=magick` (R package version 2.8.5)

Pham, L., Vu, T. H., & Tran, T. A. (2021). Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 4513–4519). doi: https://doi.org/10.1109/ICPR48806.2021.9411919

Wei, T. C., Sheikh, U., & Ab Rahman, A. A.-H. (2018). Improved optical character recognition with deep neural network. In *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)* (pp. 245–249). doi: https://doi.org/10.1109/CSPA.2018.8368720

Wyman, A., & Zhang, Z. (2023). API face value: Evaluating the current status and potential of emotion detection software in emotional deficit interventions. *Journal of Behavioral Data Science*, *3*(1), 59–69. doi: https://doi.org/10.35566/jbds/v3n1/wyman

Wyman, A., & Zhang, Z. (2025). A tutorial on the use of artificial intelligence tools for facial emotion recognition in R. *Multivariate Behavioral Research*, *60*(3), 641–655. doi: https://doi.org/10.1080/00273171.2025.2455497

Yang, K., Wang, C., Sarsenbayeva, Z., Tag, B., Dingler, T., Wadley, G., & Goncalves, J. (2021). Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer*, *37*(6), 1447–1466. doi: https://doi.org/10.1007/s00371-020-01881-x