Explaining black-box models for facial-expression detection using psychophysical reverse correlation

Rudradeep Guha, Sarra Zaied, Catherine Soladié, Pablo Arias-Sarah, Jean-Julien Aucouturier

Explaining black-box models for facial-expression detection using psychophysical reverse correlation

PREPRINT

Rudradeep Guha[1*], **Sarra Zaied**[2], **Catherine Soladié**[2], **Pablo Arias-Sarah**[3], and **Jean-Julien Aucouturier**[1]

[1]Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000 Besançon, France
[2]Department of Image Processing, IETR Institute (CNRS/CentraleSupélec/INSA Rennes/Nantes Université/Université de Rennes), Rennes, France
[3]Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LPNC, Grenoble, France

Open-source machine-learning models for automatic extraction of facial expressions have transformed psychological research, yet their black-box nature renders them opaque in terms of the features used. This is concerning because the mechanisms underlying model outputs is often as important as the model's accuracy. Here, we combined a novel face-transformation algorithm with psychophysical reverse correlation to create an explainable-AI technique. This let us uncover the features used to detect facial action units by Py-feat and Openface, two popular models for tracking facial expressions. Taking the example of smiles, we found that both models critically lack modularity: eye-region features influence model estimates of smiles, and vice-versa. Our method also helped create adversarial examples to show that task-irrelevant facial landmarks can modulate smile estimates, suggesting that the models exploit statistical shortcuts rather than anatomically meaningful features. While we found no evidence of systematic gender or ethnicity bias, we identified discrepancies between the observed and prototypical action-unit composition of emotional expressions. These findings highlight that the use of black-box models in psychological research can mask poor construct validity with high benchmark scores. However, explainability tools can counter this issue by providing mechanistic insight into how models process information, which, in addition to being good scientific practice, can help elucidate human face processing.

Keywords: reverse correlation, explainable AI, action units

1 Introduction

Machine-learning models have become ubiquitous and increasingly indispensable tools across affective computing, psychology, and cognitive science, enabling the automated extraction of multimodal behavioural signals. They are now routinely applied to facial expression recognition [1, 2], emotion recognition [3], synthesis of facial expressions for experimental stimuli [4, 5, 6], studying social interactions in psychological research [7] and for automated assessment of clinical conditions like depression [8, 9]. This ubiquity has been enabled by the accessibility and ease-of-use of toolboxes like Py-feat [10] and Openface [11]. However, studies using such tools often do not perform any form of quality checks against validated ground truth [12], and reliance on them in the scientific community raises serious concerns given the black-box nature of many machine-learning models. This means that even if a model achieves good benchmarking performance on a task, because its inner mechanisms are opaque, it is difficult to conclude actual competence at the task [13]. Models may converge on statistical regularities in training data that just happen to be predictive, rather than the theoretically meaningful constructs they are supposed to consider [14, 15, 16] (a phenomenon referred to as shortcut learning [17]). For instance, a model tasked with detecting pneumonia from radiographs did so based on spurious differences in the text embedded in the radiographs rather than any underlying medical pathology [18]. While benchmark performance is the gold standard for validating models in computer science, psychological and cognitive science research is often more concerned with *what* a model actually measures (i.e. its construct validity) than *how well* it is measured [19]. In these fields, modularity, in the sense of *information encapsulation* [20], is critical – a system estimating a given construct should be insulated from any external information that is not necessary for estimating that construct.

Issues arising from the black-box nature of machine-learning models has been well-documented, especially in terms of bias in facial analysis systems. Studies show large variance in the performance of gender classification models depending on gender and ethnicity [21], bias against under-represented groups in the data, mainly black and Asian females [22], and faster degradation of precision for black faces compared to white faces in automatic face recognition [23]. Widely used software like Google Cloud Vision, Microsoft Azure Computer Vision and Amazon Rekognition have also been found to display gender bias in terms of the kind of images used to represent women and how they are labelled and categorized, with images of women containing significantly more annotations based on physical appearance than images of men [24].

1.1 Explainable AI

Explainability in AI is an attempt to describe machine-learning systems in a manner that explicitly reveal which informational features are used for making predictions and highlight potential biases. Recently, two approaches to explaining the outputs of AI models have become prominent: sensitivity analysis [25], which quantifies the amount of change in each feature needed to affect model predictions, and layer-wise relevance propagation (LRP; [26]), which evaluates how much each feature contributes to the prediction.

*rudradeep.guha@proton.me

Explanations generated by these approaches are usually visualized with saliency maps, which highlight parts of the input that are discriminative with respect to a given class and, in so doing, provide a correlational mapping between the input and the model's prediction. For instance, LRP has been used to obtain neurophysiologically interpretable explanations for the classification of single-trial EEG by indicating the relevance of each point in high-dimensional spatio-temporal EEG data [27].

While techniques like LRP, among others like gradients [25] and excitation backprop [28], backpropagate the relevance score of features through the layers of the neural network from output to input, others propose to learn the weights of features by perturbing the inputs using, for instance, occlusion [29] and observing the corresponding effect it has on model output. One such algorithm called RISE (Randomized Input Sampling for Explanation; [30]) offers a generalizable approach by not requiring access to the model's internal features or parameters. Instead, it probes the model by randomly sub-sampling inputs and recording the resulting model outputs before generating a final saliency map, which is a linear combination of the random masks weighted by the corresponding output probabilities. Utilizing similar principles involving randomized inputs (but with additive noise instead of occlusion), recent studies have begun to explore reverse correlation, a technique originally developed for psychological research [31] to generate explanations of AI models. For instance, using a reverse-correlation procedure, random perturbations of either the texture or shape of faces revealed a strong bias towards image texture compared to image shape in a CNN trained for facial recognition [32]. Similar methodologies have also been used to demonstrate that a CNN tasked with distinguishing different facial identities places greater emphasis on the eyes, eyebrows and central face region [33] and to highlight the discrepancies between the features used by DNNs and humans to process facial identities [34]. In a comprehensive account demonstrating its versatility [35], reverse correlation is used to extract both discriminative and representative features of different kinds of classifiers across various tasks, namely, the classification of written numbers by a CNN, distinguishing speech from music with an SVM and classifying sleep stages from neurophysiological recordings.

1.2 Reverse Correlation with AU Detection Models

In this study, we leverage reverse correlation to explain facial expression recognition models by uncovering the features they use to detect facial action units in images. Action units (AUs) are basic facial expressions characterized by specific, anatomically distinct muscle movements such that activation of one muscle group does not necessarily recruit others [36]. This presumed orthogonality and independence (or modularity) of each AU, allows it to be combined to describe more complex facial expressions of emotions, an assumption we discuss critically in Section 4. By categorizing discrete facial movements into AUs, a difficult-to-quantify high-dimensional feature set is transformed into a smaller set of high-level, cognitively meaningful features. While the detection of AUs used to require manual hand-coding by experts trained in the FACS (Facial Action Unit Coding Scheme; [36]), this obstacle was overcome by machine-learning-based AU detection models whose ease-of-use and accessibility allowed widespread application in affective computing (e.g. facial expression recognition [1, 2], emotion recognition [37, 3] and detecting deepfakes [38, 39]), psychological research (e.g. facial mimicry [7] and social synchrony [40, 41]) and clinical studies (e.g. detecting pain [42], depression [43], dementia [44], Parkinson's disease [45] and atypical expressions in autism spectrum disorder [46]). Tasked with learning relevant information in the input data for maximizing classification accuracy, these AU detection models boast high accuracy. Yet, it is unclear how and why a model classifies a facial expression in a certain way. Though studies have explained emotion-recognition models [47, 48], relatively fewer studies have investigated models that detect lower-level facial features like AUs, despite some models demonstrating shortcut learning by, for example, exploiting the limited number of facial identities in AU benchmarking datasets as a shortcut for AU labels rather than learning the underlying facial muscle movements [49].

Reverse-correlation approaches to AI explainability offer the advantage of producing data representations (kernels) that are directly interpretable for psychological research, as well as direct ways to enforce these representations to new stimuli. However, the current methodologies to randomize facial stimuli for human and machine reverse-correlation experiments remain poorly suited to our research question. First, traditional reverse correlation randomizes stimuli in pixel-space [50], which means the obtained kernels are quantified relative to a single baseline stimuli (e.g. how much larger the eyes, or thinner the chin) and their morphological features cannot be directly compared to representations obtained from manipulations of other photographs. Second, while recent data-driven studies have used more advanced generative models in which reverse-correlation noise is added to the space of facial action-unit activations [34], these required using synthetic avatar faces which do not easily allow exploring possible bias across a wide variety of ethnicities or face morphologies. To directly compare reverse-correlation representations for AU-detection models across ethnicities, we would need the best of both worlds, i.e. facial transformations that can not only manipulate high-level visual features such as facial AUs in a manner that's comparable between stimuli, but are also able to do so realistically on a variety of real photographs.

To do so, we developed a novel stimulus generation technique inspired by our recent work in smile synthesis [51], which can apply random perturbations of facial expressions in images (Figure 1). We used this methodology to create thousands of random deformations on faces of different genders and ethnicities, and forced two commonly-used AU detection models – Py-feat [10] and Openface [11] – to score the deformed images as more or less likely to represent a given AU. We then use reverse correlation as an AI explainability technique [35] to extract representational features of the model for each AU, i.e. the template against which the model compares a given configuration of facial features and judges it to be a specific AU. Importantly, because these templates (referred to as *kernels* in the reverse-correlation literature; [52]) encode granular image-independent spatial information, they can be applied to any image containing a face and thus function as generative models. This generative ability allows the kernels to be visualized directly and enables them to be fed back into the

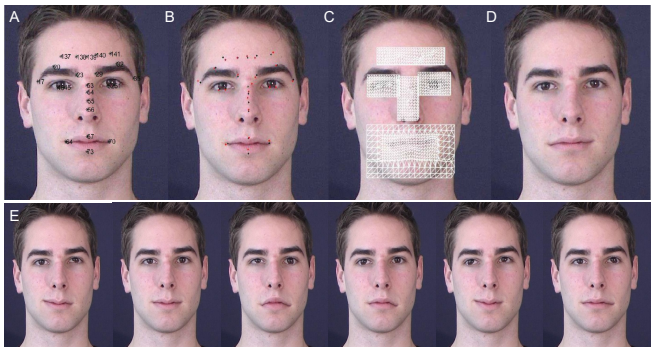

Figure 1: **Illustration of stimulus generation for AU explainability.** We generated facial reverse-correlation stimuli using a face deformation technique able to apply random perturbations of facial expressions in arbitrary face photographs. **(A)** We first use video tracking software to extract the 2D coordinates of facial landmarks on the actor's eyes, forehead, nose and mouth. While this figure illustrates the procedure with 23 landmarks from the OpenFace video tracking software, the technique used in this study instead uses 468 landmarks from the MediaPipe software **(B)** We then generate new random positions (red) for each of the landmarks by adding gaussian noise to the (x, y) coordinates of original landmarks (black). **(C)** We then deform the original photograph in the neighborhood of each landmark, using a pixel mapping technique called rigid Moving Least Squares (MLS). **(D)** The resulting image is a smooth interpolation of pixels in between landmarks, mimicking a random facial expression. **(E)** Illustration of random manipulations obtained with this procedure.

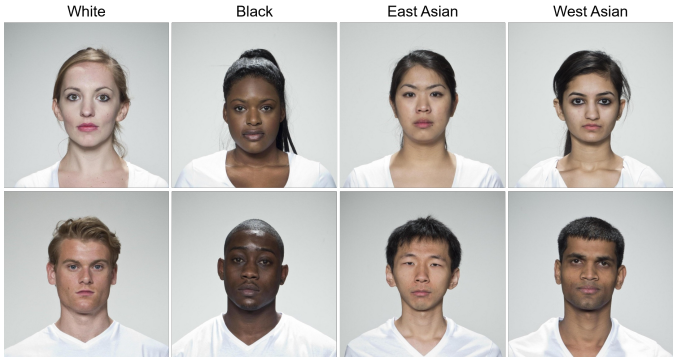

Figure 2: **Examples of facial identities from the 4 different ethnicities.** 128 such images were collected from the Chicago Faces Database [53] and the London Set from the Face Research Lab [54].

AU detection models – a process upon which all subsequent analyses are based.

Armed with these generative kernels, the representative features of AUs can then be compared to see if they are modular in image space. In other words, we ask: does the model only 'look at' the relevant facial features for classification, or does it, for instance, consider features around the mouth to classify an eye-related AU? We further explore this modularity using adversarial examples and compare the templates across different genders and ethnicities to uncover any bias. Finally, we use these templates to compute the contributions of individual AUs to make up basic emotional expressions and compare how these compositions in AU detection models differ from the theoretical composition as indicated in the FACS. While this study is concerned with Py-feat and Openface, the proposed methodology can be applied to any model.

2 Methods

2.1 Datasets

We obtained 128 high-resolution images of faces with neutral expressions from the Chicago Faces Database [53] and the London Set from the Face Research Lab [54]. Images were selected such that they were balanced across genders (male, female; 2 x 64) and the four ethnic groups defined for this study (White, Black, West Asian, East Asian; 4 x 32). Examples of identities used in the study are given in Figure 2.

2.2 Reverse-correlation stimuli

We used a custom face-transformation technique to create 2000 randomly deformed images for each image in our dataset. Using a single photograph of one actor's resting face, we first used the video tracking software Mediapipe [55] to extract the 2D coordinates of 468 facial landmarks on the actor's face (unlike the 23 visualized in Figure 1). We then generated new random positions for the landmarks by adding Gaussian noise to each of the (x, y) coordinates: $X_i^d = X_i + \mathcal{N}_i(0, \sigma)$ where $X_i = (x_i, y_i)$ are the original coordinates of the i^{th} landmark, X_i^d their newly-obtained deformed coordinates and $\mathcal{N}_i$ is a random sample from a truncated Gaussian distribution of mean $\mu = 0$, standard deviation σ, truncated at $\pm 2\sigma$. In order to obtain realistic deformations, we set σ empirically to be equal to $1/20^{th}$ of the pixel distance between the eyes of the original photograph (Figure 1B). We then deformed the original photograph to match the position of the new landmarks, using a pixel mapping technique called rigid Moving Least Squares or MLS [56]. MLS produces a function f that maps pixels in the non-deformed image to the deformed image, in a manner which transforms landmarks X_i to X_i^d, and smoothly interpolates all pixels in between (Figure 1C). We then fill the resulting triangles using affine warping (Figure 1D). Figure 1E illustrates some possible outputs of the procedure. The procedure was adapted from previous work by [51] and is provided in the FaceWarp module of CLEESE, an open-source reverse-correlation toolbox [57]. Using this procedure, a total of 265,000 images (128 facial identities $\times$ 2000 randomly deformed variants) were generated and used as stimuli for reverse correlation.

2.3 Procedure

Reverse correlation was then performed with Py-feat's XGBoost Classifier [10] and Openface [11], for each facial identity separately. A total of 2000 randomly deformed images of a facial identity were obtained using the procedure above, and grouped into random pairs to get 1000 trials. They were then

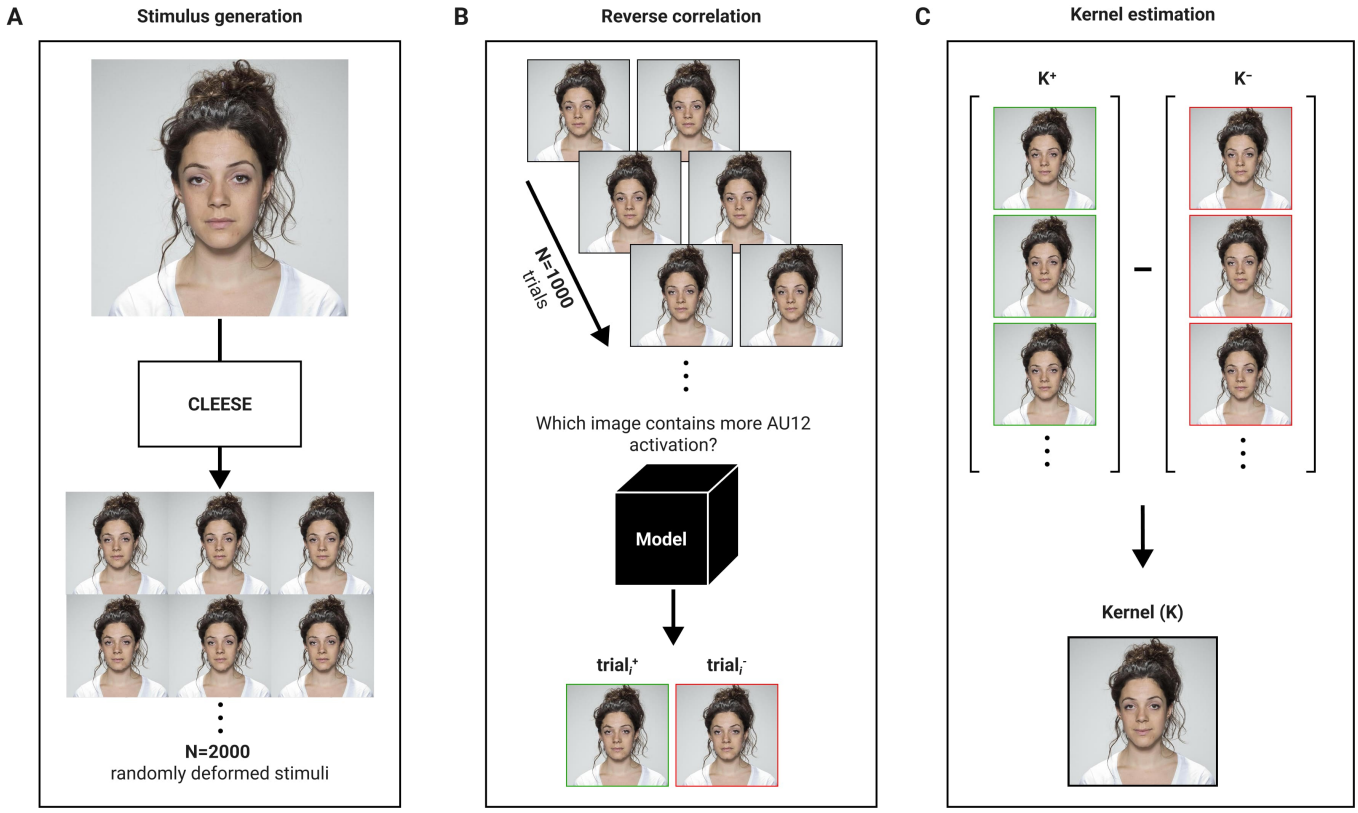

Figure 3: **The reverse correlation procedure for explaining AU detection models. (A)** Using the procedure described in the previous section, we create 2000 randomly deformed images for each of the 128 facial identities. **(B)** The reverse correlation procedure consists of assembling the randomly deformed stimuli into pairs and feeding them into the AU detection model to obtain its ratings of AUs. To obtain a binary 'response' as in typical reverse correlation experiments, we assign a value of 0 or 1 to a stimulus in the pair depending on which has a higher AU rating. **(C)** For each facial identity, and for each AU, we subtract the stimuli with positive responses (1) by the stimuli with negative responses (0) to obtain the kernel of that particular AU and facial identity.

fed into the two AU detection models. The models assigned 'scores' to all AUs in both images in a trial and the image with a higher score for an AU was considered to be its choice or response. Reverse correlating the models' responses over all trials, we extracted a kernel for each AU of each facial identity using the classification-image technique [52]. Specifically, for each AU, we computed the average random (x, y) displacement for each of the 468 landmarks in the 1000 images recognized as more representative of the AU, and subtracted the average random displacement of the images recognized as less representative. Kernels were then normalized by dividing them by the absolute sum of their values. For each facial identity, this procedure resulted in a 2×468 vector of (x, y) coordinates, representing the displacement to be applied to a given image in order to increase the probability of the resulting face being selected as more representative of a given AU (Figure 3).

3 Results

3.1 AU detection models do not enforce cognitive modularity

The normalized kernels of all 128 facial identities were averaged to obtain a single kernel for each AU. For each landmark in a kernel, we conducted one-sample t-tests against 0 (Bonferroni corrected across landmarks) on the deformation values along the x- and y-axis, such that only landmarks showing statistically significant deformations along either the x- or y-axis (or both) were retained. This resulted in a kernel containing the landmarks and the values by which to deform them in order for the model to detect the AU represented by the kernel in a face. The obtained kernels were then applied to images of the associated facial identities and fed into the model to obtain its scores for all AUs, in a process henceforth referred to as *decoding*. For each AU, we decoded the kernel (K^+) and its opposite (K^-), i.e. containing the same deformation values as in K^+ but multiplied by -1. Then, t-tests were conducted between the post-decoding AU scores of K^+ and K^- (over all facial identities), with the idea being that if the kernel did capture representative information about an AU,

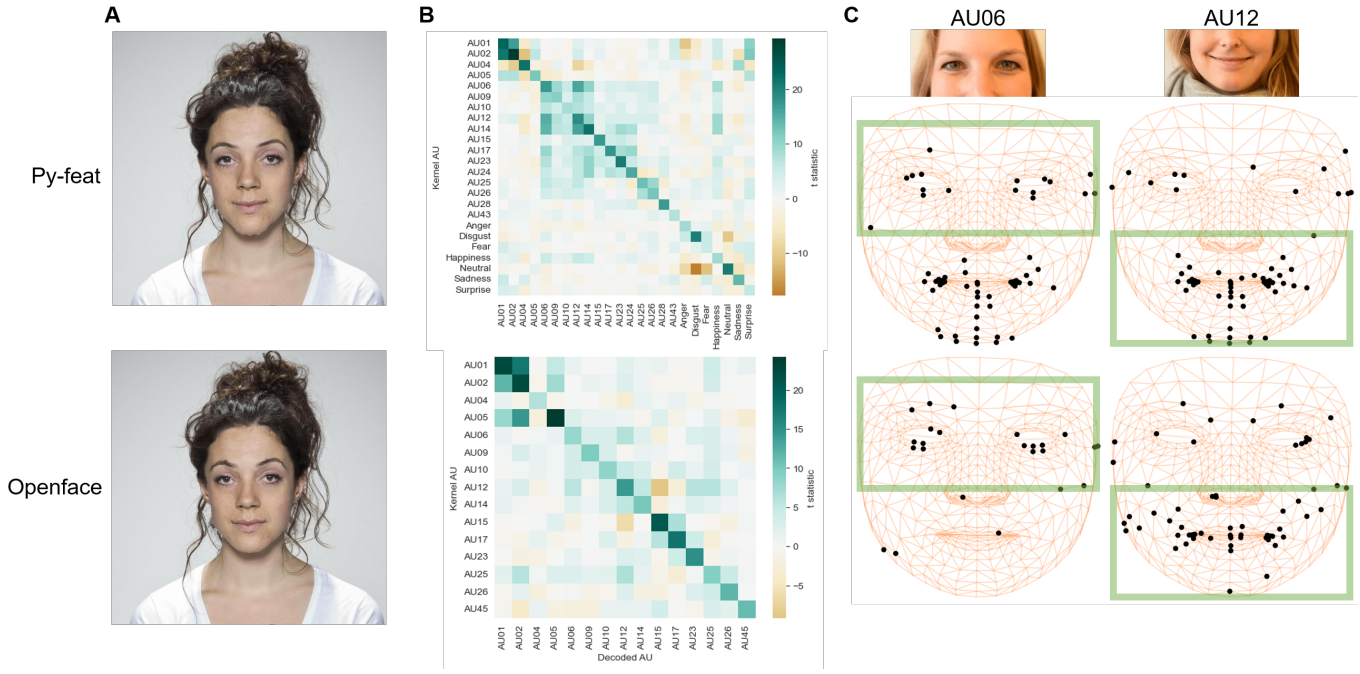

Figure 4: **Reverse correlation kernels show lack of modularity in AU detection. (A)** The AU12 (smile) reverse correlation kernels of Py-feat and Openface, visualized by applying them on a face, provide qualitative evidence for the convergence of kernels towards meaningful internal representations of the black-box models. **(B)** For most AUs, there are significant differences between model ratings of that AU when the kernel of that AU is applied as compared to when its opposite (or K^-) is applied. This suggests that the reverse correlation kernels can capture representative information about AUs, i.e. *which* specific facial features need to change and *how* for an expression to be classified as a particular AU. However, Py-feat's ratings of AU06 and AU12 appear to be correlated with each other, meaning that applying AU06 on a face changes Py-feat's rating of AU12 and vice versa. **(C)** This lack of modularity is confirmed by showing that Py-feat uses virtually the same landmarks to detect both AU06 and AU12 despite representing movements in different face regions (eyes and mouth, respectively). While Openface shows the expected pattern of results for AU06 by using mostly eye-related landmarks, it too appears to take advantage of eye-related landmarks to detect AU12.

the difference between K^+ and K^- decoding scores would be large and consequently yield a higher T statistic.

We first visualized the AU12 (smile) kernels of both Py-feat and Openface (Figure 4A) to obtain a qualitative picture of whether the reverse-correlation procedure was able to converge towards meaningful kernels. Quantitatively, we found that the reverse-correlation kernels of both Py-feat and Openface captured representative information about most AUs, as evident in the relatively large values of the T-statistic (dark green cells) along the diagonals of both heatmaps in Figure 4B. Essentially, this pattern implies that when the K^+ and K^- of a specific AU were decoded, only the ratings of that particular AU changed significantly while the others remained relatively constant. Interestingly though, we observed notable violations of this pattern, especially in Py-feat with AU06 (cheek raiser), AU12 (lip corner puller) and AU14 (dimpler), suggesting that a change in the intensity of AU06, for instance, induced a significant change in Py-feat's score of AU12.

Concentrating on the two most non-modular AUs, we investigated which landmarks were significantly important for detecting AU06 and AU12 in Py-feat and Openface (Figure 4C). In the case of AU06, we found that Py-feat, in addition to expected landmarks in the eye-region of the face, unexpectedly used a large number of landmarks around the mouth, while Openface mostly only used landmarks around the eyes. Remarkably, the landmarks used by Py-feat to detect AU06 were almost identical to those used to detect AU12. Contrary to expectations, both Py-feat and Openface demonstrated significant involvement of landmarks in the eye-region in addition to the mouth-region for the detection of AU12.

3.2 Kernels can be used to generate adversarial examples

To further investigate the influence of 'task-irrelevant' landmarks on AU ratings, we subjected the AU detection models to adversarial examples. From Py-feat's AU12 kernel, we isolated 3 landmarks: 280 (cheek), 285 (inner eyebrow) and 373 (lower left eye). These landmarks were chosen because they were found to significantly influence AU12 ratings in both Py-feat and Openface and, by virtue of not being explicitly mouth-related landmarks, were deemed less relevant for detecting the mouth-related AU12. Thus, we created 3 kernels containing a single landmark and the corresponding deformations along the x- and y-axis as specified by the AU12 kernel.

First, we separately applied each of the 3 single-landmark kernels to a neutral face for different values of the scaling factor s

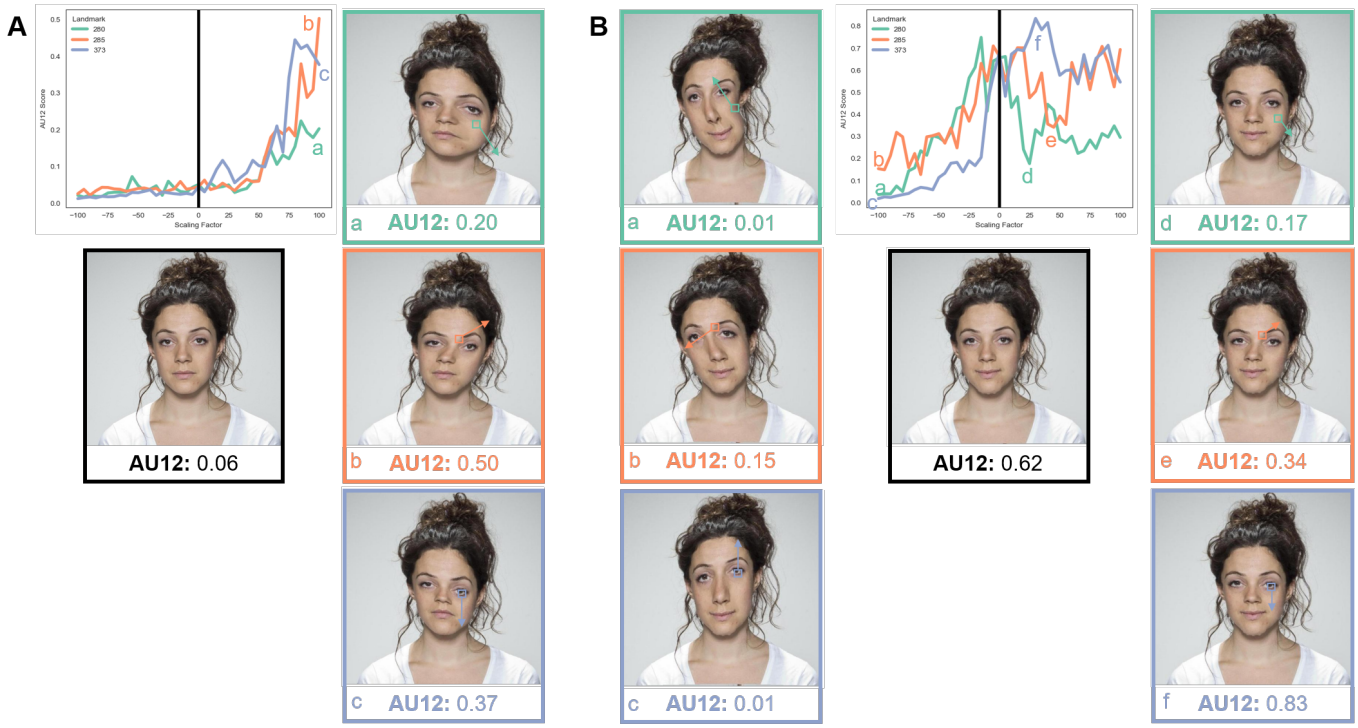

Figure 5: **Adversarial examples on Py-feat.** We isolated 3 landmarks in Py-feat's AU12 kernel found to have a significant impact on AU12 ratings despite not being in the mouth region - namely, 280 (left cheek, green), 285 (left inner eyebrow, orange) and 373 (lower left eye, blue). **(A)** On a neutral face, individual landmarks are deformed by varying magnitudes (i.e. by a scaling factor s, where $s \in [-100, -90, ..., 90, 100]$ with negative values representing deformation in the opposite direction). Deforming the landmarks by large positive values increases Py-feat's ratings of AU12 **(a, b, c)**, but the same is not observed for negative values of s. **(B)** A similar procedure is conducted on a face upon which the average AU12 kernel has been applied (i.e. a face already containing AU12). Large deformations by negative values of s dramatically decrease Py-feat's ratings of AU12 **(a, b, c)** despite changes occurring largely in the general shape of the head while leaving the smiles intact. Deforming landmarks by increasingly large positive values of s demonstrates significant variability in Py-feat's AU12 ratings with relatively small positive values producing large changes in AU12 ratings **(d, e, f)** despite, in most cases, very little discernible effects on the face.

($s \in [-100, -90, ..., 90, 100]$), i.e. the value by which to multiply the deformation values. None of the 3 landmarks affected AU12 ratings when their deformation values were scaled by negative values of s. However, with $s = 100$, Py-feat's AU12 ratings increased from 0.06 for the original neutral face to 0.2, 0.5 and 0.37 for landmarks 280, 285 and 373, respectively (Figure 5A). The same procedure was repeated with the base figure containing a decoded AU12 kernel (i.e. already containing a smile). Large negative values of s dramatically decreased Py-feat's AU12 ratings of the faces from 0.62 to 0.01 for landmarks 280 and 373, and to 0.15 for landmark 285 (Figure 5Ba, b and c). On the other hand, positive values of s resulted in a more volatile trend without any consistent increase or decrease in AU12 ratings. Even relatively small positive values of s that did not produce any discernible changes to the face resulted in a sharp decrease in AU12 rating from 0.62 to 0.17 for landmark 280 ($s = 25$) and an increase to 0.83 for landmark 373 ($s = 30$), while the more visible deformation of landmark 285 ($s = 45$) caused the AU12 rating to fall by half (Figure 5Bd, e and f).

3.3 AU detection models do not show strong gender or ethnicity bias

To investigate whether a model's internal representations of AUs vary with gender and/or ethnicity, we compared the kernels of the two groups of interest (e.g. male vs. female). We conducted t-tests between the deformation values of each of the 468 landmarks (Bonferroni corrected) over all faces in the two groups. This was done for every available AU in the model and visualized in the form of a heatmap showing, for each AU, whether there was at least one landmark with deformation values that were significantly different between the two groups being compared. Results showed that the number of AUs with at least one significantly different landmark was greater in Openface as compared to Py-feat. Compared to eye-related AUs, there were also almost twice as many mouth-related AUs with at least one significantly different landmark. Globally, the sparsity of the distribution of significant AUs and the lack of consistency across group comparisons in terms of which AUs were found to be significant did not indicate the presence of any systematic bias against a particular gender or ethnicity. Despite the total number of AUs/emotions with differences between groups, closer inspection revealed that

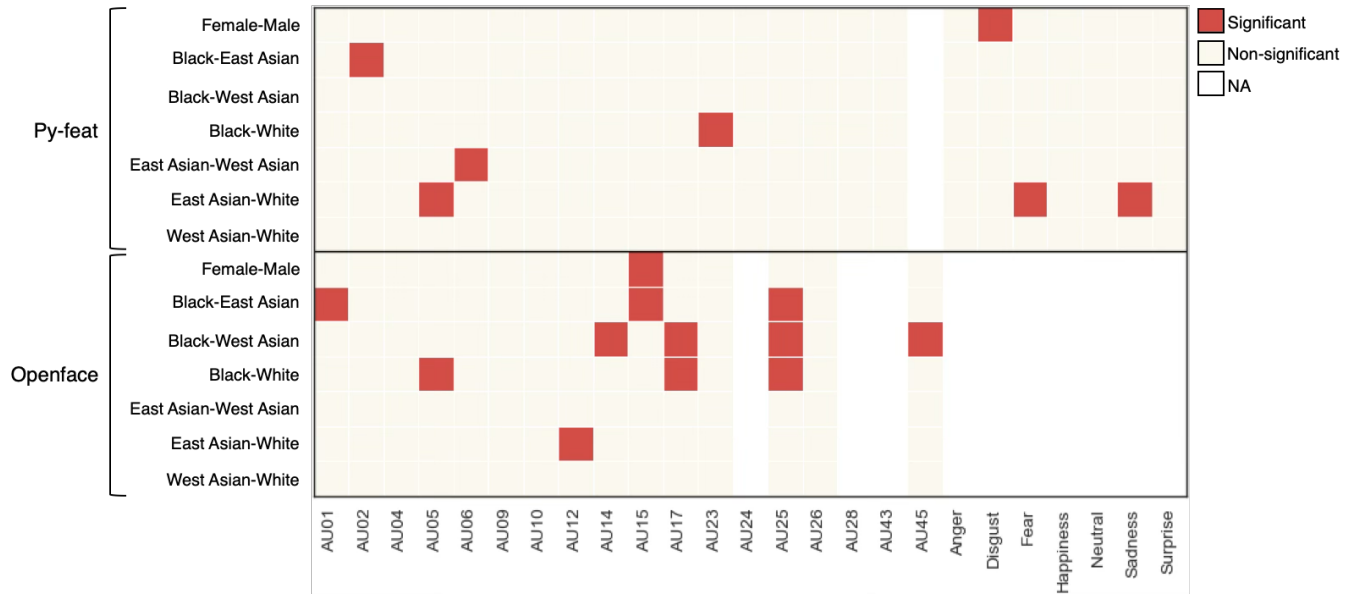

Figure 6: **Lack of systematic gender and ethnicity bias in Py-feat and Openface.** Conducting t-tests between deformation values of the kernels between groups reveals that no specific group comparison consistently yields a greater number of significantly different AUs (as seen in the sparse and inconsistent distribution of red cells in the heatmap). Examining significant AUs more closely shows that the significant differences are driven only by a few landmarks (2 on average and a maximum of 4 in one comparison for one AU). On the whole, the relatively similar spatial distribution of landmarks used by the models for different groups suggests that these models do not show systematic bias against any specific gender or ethnicity.

most of these comparisons contained only one or two significantly different landmarks, with only a single comparison between West Asian and Black identities on AU45 showing a maximum of 4 (Figure 6).

3.4 Kernels can quantify how AI models compose emotions in terms of AUs

Finally, we took the decoding scores of Py-feat's emotion kernels for all facial identities and used them to investigate the model's composition of basic emotions (happiness, sadness, surprise, fear, anger and disgust; [58]) in terms of AUs. We used ridge regression with 5-fold cross-validation to model the relationship between standardized decoding scores of AUs and the decoding score of the corresponding emotional expressions. Regularization parameters were optimized across a grid from 10^{-3} to 10^{3}, and coefficients were extracted to quantify each AU's contribution to emotion prediction while controlling for multicollinearity among AUs. The coefficients of the independent variables were taken to represent the extent of their contributions to the detection of the emotion specified as the dependent variable and visualized using radar plots. These plots were then compared qualitatively with the theoretical composition of emotions, i.e. the AUs whose combination is *supposed* to give rise to a given emotion as per the FACS.

Results showed that while the relationship between AUs and most of the emotions was relatively well-matched with theoretical expectations, there were a few discrepancies (Figure 7). While the composition of happiness showed the expected contributions of AU06 and AU12, it also showed a large contribution from AU02 (outer brow raiser). Similarly, both sadness and surprise demonstrated largely the same patterns as their theoretical compositions, but with the incongruous addition of AU25 (lips part) [36, 59]. The composition of fear displayed lower contribution from AU04 (brow lowerer) and greater contribution from AU26 (jaw drop) than expected. On the other hand, the composition of anger and disgust deviated significantly from expectations, with anger showing contributions from a wide range of other AUs while disgust showed the greatest contribution from AU06, AU09 (nose wrinkler) and AU12 rather than the expected combination of AU09 and AU15 (lip corner depressor). This corresponds with studies showing that negative emotions like anger and disgust show greater variability across individuals and cultures than positive emotions like happiness [60].

4 Discussion

In this study, we leverage techniques from psychophysics, namely reverse correlation, to explain machine-learning action-unit detection models. We show that reverse-correlation kernels capture representative information about the features utilized by Py-feat and Openface to make judgements about the presence of AUs. Taking advantage of this information, we first highlight the lack of modularity in these models by showing that ratings of, e.g. AU12, are influenced by features in the eye-region in both models. Second, we use information from AU12 kernels to isolate some landmarks that ideally should not impact a model's ratings and show that the models can be sensitive to small changes that are seemingly imperceptible to humans. Furthermore, we demonstrate the

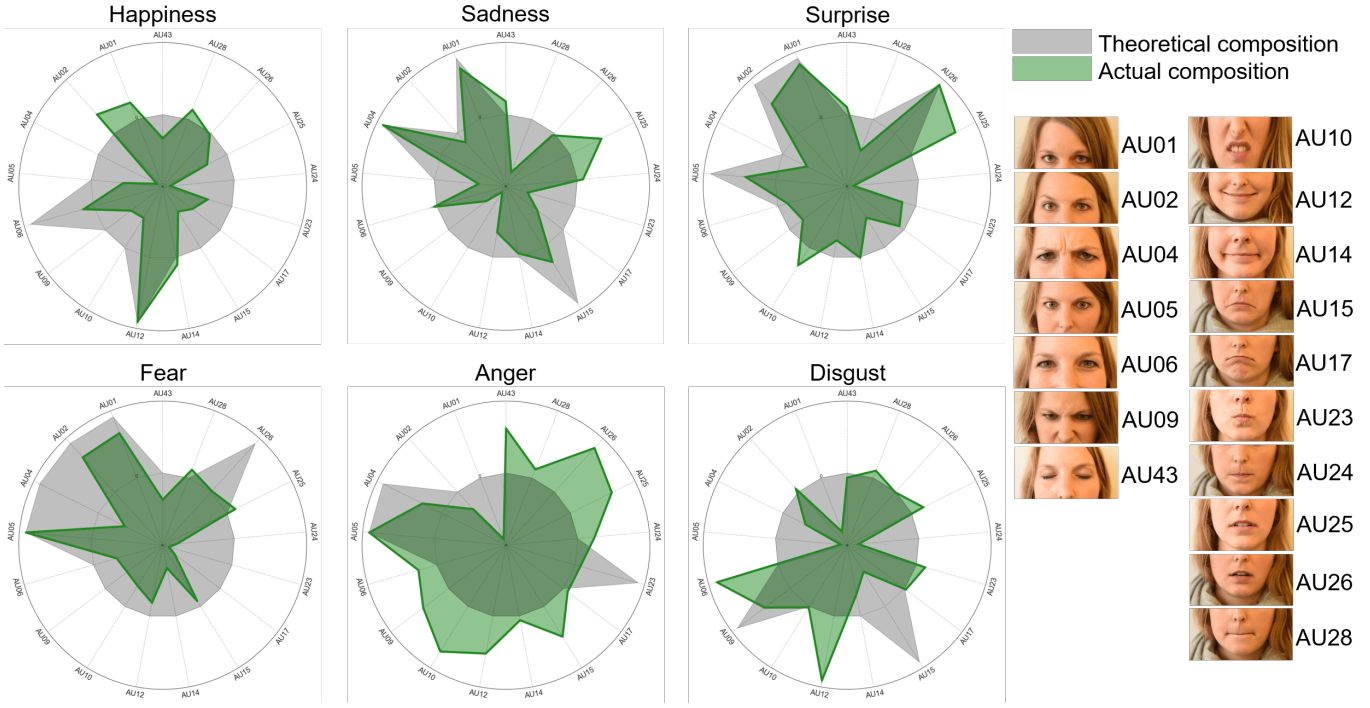

Figure 7: **Composition of emotions in terms of AUs.** Applying Py-feat's emotion kernels on the 128 facial identities, we model their emotion ratings with the corresponding ratings of all other AUs as predictors. Taking the regression coefficients (green) as the extent of contribution to the classification of a face as an emotion, they are compared against the theoretical composition of emotions (black) suggested by the FACS. Results reveal several discrepancies, like the involvement of AU01 and AU02 in happiness, AU25 in sadness and surprise and AU26 in fear. The compositions of anger and disgust, in particular, display significant deviance from the theoretical composition, with many different AUs involved in anger and AU06 and AU12 appearing to be the largest contributors to disgust.

method's potential to uncover gender or ethnicity differences in how the models internally represent facial expressions. While these differences exist, they are sparsely distributed across AUs and participant categories, and we did not find any systematic bias against one specific gender or ethnicity. Finally, we demonstrate that reverse-correlation kernels can be used to decompose high-level emotional facial expressions into their component AUs, with potential for achieving more fine-grained control over complex facial expressions and thus providing useful insight into how models internally represent these emotions in comparison to what can be expected theoretically.

The fact that kernels capture representative information about the AU detection models shows the viability of reverse correlation as a technique to explain machine-learning based models, a point also made by [35]. Unlike some other algorithms for explainability, kernels computed from the reverse correlation procedure not only reveal *which* parts of the image affect model decisions but also *how* those parts need to be manipulated to induce that effect. Access to this information expands the utility of these kernels by allowing us to apply them to the original stimuli and probe the model again, in a process referred to in this study as decoding. Decoding the kernels thus provides a mechanism for quantifying how much of the information contained within them is actually influential in driving the model's decisions. In further work, one could also potentially present the model's kernels as image stimuli to human participants, as well as participants' kernels as stimuli to the machine-learning models, and examine the differences between the features used by machines and humans.

4.1 Modularity

The finding that Py-feat's AU06 kernel focuses heavily on landmarks around the mouth instead of the eyes and that the AU12 kernel focuses on the eyes in addition to the mouth shows that AU detection models do not always ensure modularity. This is significant because the FACS aims to break down complex social signals into combinations of independent, smaller units. If these smaller units are conflated with each other by AU detection models or are not treated modularly, AU activity inferred by machine-learning models is potentially confounded. For instance, using Py-feat to detect 'genuine' or Duchenne smiles, which are supposed to be a combination of AU06 and AU12, will likely produce incorrect results given the positive dependence of the two AUs in Py-feat and its use of virtually the same landmarks to detect both (Figure 4C). The finding that both Py-feat and Openface detect AU12 using landmarks around the eyes in addition to those around the mouth poses similar problems. One possible cause for the lack of modularity could be the fact that they were trained on datasets containing naturalistic expressions. If the expres-

sions were annotated for each AU independently and the distribution of the AUs is not independent, then the activity of one AU could be used as a shortcut to classify the other [17]. Consequently, a model trained on a dataset featuring faces that only display parametric activation of a single AU, or with a random i.i.d. (independent and identically distributed) combination of AUs, might display more modularity. It is worth noting that the FACS was originally proposed as a descriptive coding system, cataloguing how muscle movements combine to form AUs and how AUs in turn combine to form emotional expressions. Modularity was not an explicit claim of the system. However, researchers have, over time, operationalized AUs as discrete, independently measurable units. The lack of modularity we observe is therefore less a refutation of the FACS than a cautionary tale about how it has come to be understood and used in the scientific community.

The issue of modularity is further highlighted by evidence showing that AU detection models are sensitive to 'adversarial examples' featuring changes in ostensibly irrelevant regions of a face. We show that manipulating individual landmarks on the cheek, inner eyebrow, and lower left eye of a neutral face induces a large increase in Py-feat's rating of AU12 despite the lip corners actually moving downwards in some cases. Similarly, we find that manipulating these landmarks on an already smiling face also dramatically reduces or increases AU12 ratings depending on the direction of the manipulation. A few of these manipulations produced barely perceptible differences in the face, and yet yielded a significant change (decrease for the landmark on the cheek and increase for the landmark on the lower left eye) in the model's AU12 rating. Sensitivity to these seemingly 'task-irrelevant' landmarks could reveal analytical strategies taken by the models to quantify task-relevant features. For instance, it is possible that faced with the task of learning a linear combination of landmark positions that correlates with AU12, models converge on estimating the distance between the lip corner and, e.g. the ipsilateral upper eyelid. Deformation of the eyelid could then lead the model to wrongly attribute it to a smiling expression. However, given that large deformations tend to create unnatural modifications to facial morphology, it might be anticipated that Py-feat, having been trained on naturalistic facial expressions, would fail to detect AU12 in images featuring unnatural face shapes. Although this could account for the reduced AU12 ratings observed when such deformations are applied to smiling faces, the persistence of elevated AU12 ratings when large deformations are applied to neutral faces remains noteworthy. This pattern suggests that, beyond global changes in face morphology, Py-feat also shows sensitivity to specific features being manipulated.

4.2 Bias

While there were examples of differences in AU detection across participant groups, we could not draw any conclusions about the presence of bias against a particular gender or ethnicity. For instance, while Openface displayed relatively consistent differences for AU15, AU17 and AU25 across group comparisons, closer examination revealed that only 2-3 landmarks actually differed in terms of how they needed to be deformed. The lack of large-scale differences in facial features suggests that the machine-learning models tested here do not exhibit the kind of social-cognitive stereotypes usually encountered in human participants (e.g. systematically associating one ethnicity with one personality trait [61]). However, the presence of small-scale differences in a handful of landmarks might instead suggest a more granular algorithmic bias. This is borne out by the adversarial examples showing large changes in Py-feat's AU12 ratings as a result of deforming individual landmarks. This could potentially be investigated by comparing the effects of the same adversarial examples on different genders and ethnicities. It is also possible that the differences are the result of differing face morphology, depending on the gender or ethnicity. Differences in face morphology driving differential ratings of AUs and emotions raise the question of what constitutes bias. Is a model biased if its outputs are different for different groups? Or is it biased if it disregards differences and generates the same outputs for different groups? A recent study found that their seemingly effective algorithm to mitigate biased outputs in LLMs failed when the LLM was deployed in a different context [62]. This context-dependence highlights the need for a more dynamic understanding of bias instead of ascribing it statically. These two competing notions of algorithmic fairness – statistical parity (differential outputs across groups constitutes bias) and calibration (same outputs across groups constitutes bias) – are often at odds with each other and have been shown to be mathematically incompatible [63].

4.3 Composition of emotions in terms of AUs

Discrepancies in the composition of emotions (in terms of AUs) in this study and the theoretical composition described by the FACS and the classical affective-science literature, if taken at face value, suggest issues in Py-feat's representations of the emotions, particularly for anger and disgust. However, studies have shown that while machine-learning models show high accuracies for classifying emotions in standardized datasets containing trained actors displaying prototypical posed emotions, their performance is more variable for emotional displays of non-standardized spontaneous emotional expressions [64, 65]. Since Py-feat's models for emotion detection were trained on both posed and naturalistically elicited emotional expressions [10], it is thus possible that Py-feat captures the true composition or at least a greater amount of the variability of emotional expressions [66]. Indeed, several studies show that theoretically proposed AU patterns in emotional expressions are not backed up by empirical findings [67], with actors often either not displaying all the prototypical AUs or displaying AUs other than those predicted [68]. Moreover, previous work also suggests that machine-learning classification of non-standardized portrayals of emotional expressions is often worse for negative emotions like anger and disgust than for happiness [69]. This could be an explanation for the greater discrepancies between empirical and theoretical compositions of anger and disgust that we observe here.

4.4 Broader implications for psychological research

First, this study is not intended as a criticism of the models. On the contrary, the field of psychology is indebted to the developers of these models for their efforts in creating these models, contributing to advancing the field, and for making them open-source. The reverse-correlation framework

demonstrated in this study is thus more of an effort to understand these models more deeply and advance performance and interpretability as complementary goals. Taken together, our results highlight the need for careful evaluation of the outputs of such black-box models. Indeed, the authors of the Py-feat toolbox acknowledge that the datasets on which their models are trained may be unbalanced and advise users to verify the outputs. However, the accessibility and ease-of-use of these tools, combined with the difficulty of manually verifying their outputs, often means that they are taken at face value. The tendency to highlight model performance on benchmarking datasets further adds to the issue by focusing on task performance statistics instead of providing more qualitative explanations of how such performance is achieved [13]. This is a more general concern in AI discourse – namely, the masking of poor construct validity of models by high performance/accuracy. Being cautious of such 'shortcut learning' or 'approximate retrieval', where models measure statistical regularities in the training data rather than the theoretically meaningful quantity they are supposed to measure, is of particular importance in research settings where experimental control, precision and interpretability are crucial.

Given that the FACS does not specify exactly how muscles need to be configured to 'make' a specific AU, it becomes difficult to use AUs as a basis for creating precise stimuli according to precise specifications. It is here that the kernels obtained from reverse correlation can come in and bridge the gap between low-level landmarks and high-level action units by mapping emotional expressions and AUs to their corresponding landmarks. As shown here with the decoding paradigm, kernels obtained from performing reverse correlation on machine-learning models can function as generative models. The ability to visualize these kernels on images allows them to be used as bases for creating images or videos of persons displaying a particular AU or emotional expression. This is particularly useful for generating synthetic media (for animators, for instance) as well as creating experimental stimuli containing fine-grained manipulations.

4.5 Limitations

One limitation of this study is a lack of diversity and naive grouping of the different ethnicities. First, our categorization of facial identities based on ethnicity was perhaps too broad (with 4 groups: White, Black, West Asian and East Asian). Additionally, our conclusions about gender and ethnicity bias were based on a relatively low number of facial identities in each group. The fact that we had a total of 128 facial identities meant that each ethnicity only contained 32 facial identities. A better estimate of the presence of any bias could be obtained by incorporating a greater number of facial identities. Moreover, given the lack of realism or ecological validity of some of the transformations in this study, future work could also utilize a more powerful class of image generation models with the ability to generate more naturalistic deformations and manipulate not only face landmarks but also other image features like texture. This would allow one to generate more complex and multi-faceted kernels and adversarial tests for the AU detection models. Finally, it should be noted that while we did not find any bias against an ethnicity or gender in these models, it remains possible that they may be *used* to discriminate against certain groups. In other words, technology may be objective in isolation, but its use by human agents can often imbue it with subjectivity through the transfer, deliberate or otherwise, of our preconceptions and biases [70, 71].

Acknowledgements

The work was funded by ANR SOUNDS4COMA and conducted in the framework of the EIPHI Graduate school (ANR-17-EURE-0002 contract). All data reported in the paper are available on request.

References

1. Tian, Y.-I., Kanade, T. & Cohn, J. F. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* **23,** 97–115 (2001).
2. Zhao, Y. *et al. Action unit driven facial expression synthesis from a single image with patch attentive GAN* in *Computer Graphics Forum* **40** (2021), 47–61.
3. Zhi, R., Zhou, C., Li, T., Liu, S. & Jin, Y. Action unit analysis enhanced facial expression recognition by deep neural network evolution. *Neurocomputing* **425,** 135–148 (2021).
4. Roesch, E. B. *et al.* FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal of Nonverbal Behavior* **35,** 1–16 (2011).
5. Fan, Y., Li, V. O. & Lam, J. C. Facial expression recognition with deeply-supervised attention network. *IEEE transactions on affective computing* **13,** 1057–1071 (2020).
6. Guo, Y. *et al. StyleAU: StyleGAN based Facial Action Unit Manipulation for Expression Editing* in *2023 IEEE International Joint Conference on Biometrics (IJCB)* (2023), 1–10.
7. Arias-Sarah, P. *et al.* Aligning the smiles of dating dyads causally increases attraction. *Proceedings of the National Academy of Sciences* **121,** e2400369121 (2024).
8. Ringeval, F. *et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition* in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop* (2019), 3–12.
9. De Melo, W. C., Granger, E. & Hadid, A. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE transactions on affective computing* **13,** 1581–1592 (2020).
10. Cheong, J. H. *et al.* Py-feat: Python facial expression analysis toolbox. *Affective Science* **4,** 781–796 (2023).
11. Baltrušaitis, T., Robinson, P. & Morency, L.-P. *Openface: an open source facial behavior analysis toolkit* in *2016 IEEE winter conference on applications of computer vision (WACV)* (2016), 1–10.
12. Nota, N., Trujillo, J. P. & Holler, J. Detecting facial signals in conversational corpora using OpenFace compared to manual annotation: Challenges and recommendations.

13. Firestone, C. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences* **117,** 26562–26571 (2020).
14. Arras, L., Horn, F., Montavon, G., Müller, K.-R. & Samek, W. *Explaining predictions of non-linear classifiers in NLP* in *Proceedings of the 1st Workshop on Representation Learning for NLP* (2016), 1–7.
15. Arras, L., Horn, F., Montavon, G., Müller, K.-R. & Samek, W. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one* **12,** e0181142 (2017).
16. Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R. & Samek, W. *Analyzing classifiers: Fisher vectors and deep neural networks* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 2912–2920.
17. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2,** 665–673 (2020).
18. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15,** e1002683 (2018).
19. Schyns, P. G., Bonnar, L. & Gosselin, F. Show me the features! Understanding recognition from the use of visual information. *Psychological science* **13,** 402–409 (2002).
20. Fodor, J. A. *The modularity of mind* (MIT press, 1983).
21. Buolamwini, J. & Gebru, T. *Gender shades: Intersectional accuracy disparities in commercial gender classification* in *Conference on fairness, accountability and transparency* (2018), 77–91.
22. Serna, I., Pena, A., Morales, A. & Fierrez, J. *InsideBias: Measuring bias in deep networks and application to face gender biometrics* in *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), 3720–3727.
23. Majumdar, P., Mittal, S., Singh, R. & Vatsa, M. *Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 3786–3795.
24. Schwemmer, C. *et al.* Diagnosing gender bias in image recognition systems. *Socius* **6,** 2378023120967171 (2020).
25. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
26. Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* **10,** e0130140 (2015).
27. Sturm, I., Lapuschkin, S., Samek, W. & Müller, K.-R. Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods* **274,** 141–145 (2016).
28. Zhang, J. *et al.* Top-down neural attention by excitation backprop. *International Journal of Computer Vision* **126,** 1084–1102 (2018).
29. Zeiler, M. D. & Fergus, R. *Visualizing and understanding convolutional networks* in *European conference on computer vision* (2014), 818–833.
30. Petsiuk, V., Das, A. & Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).
31. Ponsot, E., Burred, J. J., Belin, P. & Aucouturier, J.-J. Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences* **115,** 3972–3977 (2018).
32. Xu, T. *et al.* Deeper interpretability of deep networks. *arXiv preprint arXiv:1811.07807* (2018).
33. Tian, X., Song, Y. & Liu, J. Decoding face identity: A reverse-correlation approach using deep learning. *Cognition* **254,** 106008 (2025).
34. Daube, C. *et al.* Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns* **2** (2021).
35. Thoret, E., Andrillon, T., Léger, D. & Pressnitzer, D. Probing machine-learning classifiers using noise, bubbles, and reverse correlation. *Journal of neuroscience methods* **362,** 109297 (2021).
36. Ekman, P. & Friesen, W. V. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).
37. Nadeeshani, M., Jayaweera, A. & Samarasinghe, P. *Facial emotion prediction through action units and deep learning* in *2020 2nd International Conference on Advancements in Computing (ICAC)* **1** (2020), 293–298.
38. Jaleel, Q. & Hadi, I. *Facial action unit-based deepfake video detection using deep learning* in *2022 4th International Conference on Current Research in Engineering and Science Applications (ICCRESA)* (2022), 228–233.
39. Liao, X., Wang, Y., Wang, T., Hu, J. & Wu, X. FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Transactions on Circuits and Systems for Video Technology* **33,** 7236–7251 (2023).
40. Cheong, J. H., Molani, Z., Sadhukha, S. & Chang, L. J. Synchronized affect in shared experiences strengthens social connection. *Communications Biology* **6,** 1099 (2023).
41. Nota, N., Trujillo, J. P. & Holler, J. Conversational eyebrow frowns facilitate question identification: An online study using virtual avatars. *Cognitive Science* **47,** e13392 (2023).
42. Bouazizi, M., Feghoul, K., Wang, S., Yin, Y. & Ohtsuki, T. A non-invasive approach for facial action unit extraction and its application in pain detection. *Bioengineering* **12,** 195 (2025).
43. Akbar, H. *et al. Exploiting facial action unit in video for recognizing depression using metaheuristic and neural networks* in *2021 1st International conference on computer science and artificial intelligence (ICCSAI)* **1** (2021), 438–443.
44. Kolosov, D. *et al. Association of Facial Action Units with Emotions in People Living with Dementia Listening to Music* in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)* (2024), 1195–1199.
45. Razzouki, A. F. *et al.* Leveraging Action Unit Derivatives for Early-Stage Parkinson's Disease Detection. *IRBM* **46,** 100874 (2025).

46. Ji, Y. *et al.* Hugging rain man: a novel facial action units dataset for analyzing atypical facial expressions in children with autism spectrum disorder. *IEEE Transactions on Affective Computing* (2025).
47. Del Castillo Torres, G., Roig-Maimó, M. F., Mascaró-Oliver, M., Amengual-Alcover, E. & Mas-Sansó, R. Understanding how CNNs recognize facial expressions: a case study with LIME and CEM. *Sensors* **23,** 131 (2022).
48. Weitz, K., Hassan, T., Schmid, U. & Garbas, J.-U. Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen* **86,** 404–412 (2019).
49. Ning, M., Salah, A. A. & Ertugrul, I. O. Representation learning and identity adversarial training for facial behavior understanding. *arXiv preprint arXiv:2407.11243* (2024).
50. Maister, L., De Beukelaer, S., Longo, M. R. & Tsakiris, M. The self in the mind's eye: Revealing how we truly see ourselves through reverse correlation. *Psychological Science* **32,** 1965–1978 (2021).
51. Arias, P. *et al.* Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing* **11,** 507–518 (2018).
52. Murray, R. F. Classification images: A review. *Journal of vision* **11,** 2–2 (2011).
53. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* **47,** 1122–1135 (2015).
54. DeBruine, L. & Jones, B. Face research lab London set. *(No Title)* (2017).
55. Lugaresi, C. *et al.* Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
56. Schaefer, S., McPhail, T. & Warren, J. in *ACM SIGGRAPH 2006 Papers* 533–540 (2006).
57. Burred, J. J., Ponsot, E., Goupil, L., Liuni, M. & Aucouturier, J.-J. CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PloS one* **14,** e0205943 (2019).
58. Ekman, P. & Friesen, W. V. Constants across cultures in the face and emotion. *Journal of personality and social psychology* **17,** 124 (1971).
59. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* **20,** 1–68 (2019).
60. Cordaro, D. T. *et al.* Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* **18,** 75 (2018).
61. Gingras, F. *et al.* Pain in the eye of the beholder: Variations in pain visual representations as a function of face ethnicity and culture. *British Journal of Psychology* **114,** 621–637 (2023).
62. Ma, S., Salinas, A., Henderson, P. & Nyarko, J. *Breaking down bias: On the limits of generalizable pruning strategies* in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025), 2437–2450.
63. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On fairness and calibration. *Advances in neural information processing systems* **30** (2017).
64. Dupré, D., Krumhuber, E. G., Küster, D. & McKeown, G. J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *Plos one* **15,** e0231968 (2020).
65. Küntzler, T., Höfling, T. T. A. & Alpers, G. W. Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Frontiers in psychology* **12,** 627561 (2021).
66. Jack, R. E., Garrod, O. G., Yu, H., Caldara, R. & Schyns, P. G. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* **109,** 7241–7244 (2012).
67. Sato, W., Hyniewska, S., Minemoto, K. & Yoshikawa, S. Facial expressions of basic emotions in Japanese laypeople. *Frontiers in psychology* **10,** 259 (2019).
68. Gosselin, P., Kirouac, G. & Doré, F. Y. Components and recognition of facial expression in the communication of emotion by actors. *Journal of personality and social psychology* **68,** 83 (1995).
69. Stöckli, S., Schulte-Mecklenbeck, M., Borer, S. & Samson, A. C. Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior research methods* **50,** 1446–1460 (2018).
70. Fountain, J. E. *Building the virtual state: Information technology and institutional change* (Rowman & Littlefield, 2004).
71. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **54,** 1–35 (2021).