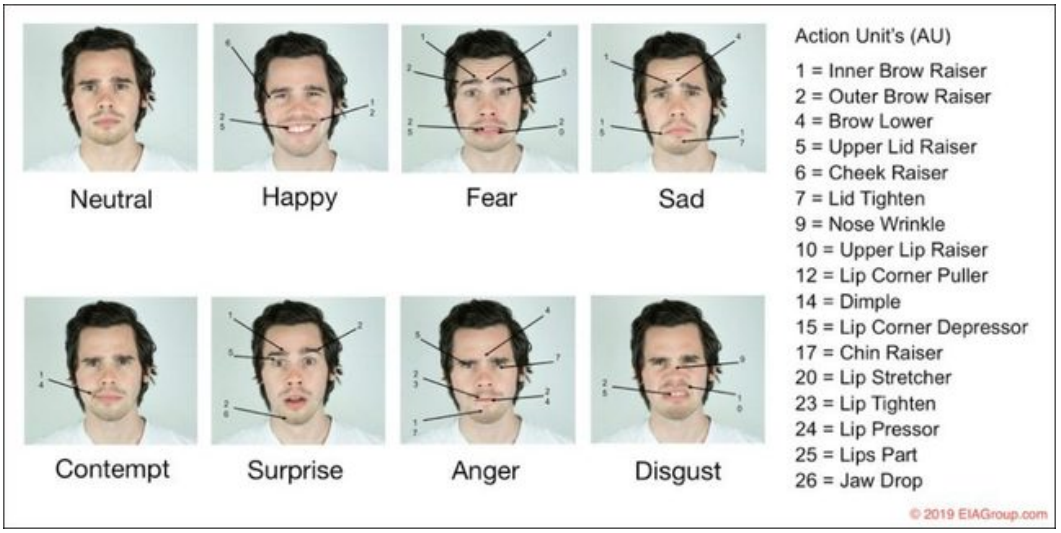
The Challenge of Using AI to Change Facial Expressions

Originally published April 07, 2023 at https://blog.metaphysic.ai/the-challenge-of-using-ai-to-change-facial-expressions/

Currently, the AI facial synthesis scene is carving out its space in fairly broad strokes, concentrating on the accurate recreation of individuals, using variations of a by-now familiar slate of technologies, including autoencoders, Generative Adversarial Networks (GANs), Neural Radiance Fields (NeRF), and latent diffusion models (LDMs).

Recreating a truly authentic likeness, or transferring a likeness in an efficient and resource-rational manner, is such a challenge for the sector, that some of the finer roadblocks have been de-prioritized, for the time being.

One of these is *expression editing* – the ability to recognize and change the facial affect tonality of a reproduced face (i.e., *angry*, *bored*, *concerned*, etc.).



Though increasingly coming under criticism, the Facial Action Coding System (FACS) is the only significant and widely-adopted attempt to date to quantify the relationship between experienced emotions and the way that they manifest in the human face. Source: https://www.eiagroup.com/the-facial-action-coding-system/

Despite notable progress in neural facial synthesis, expression editing is, arguably, in its earliest days. Anyone who uses the current version of Adobe Photoshop, for instance, and who has played around with some of the expression-editing settings in its AI-based Neural Filters, will know what a terrible job it makes of trying to control the amount of 'happiness' or 'anger' in a submitted face, or in generally trying to manipulate facial expressions in a convincing manner:

Turning 'anger' up to the maximum in the Smart Portrait > Expressions section of Photoshop's Neural Filters usually obtains an unsatisfactory result, because the facial manifestation of that emotion is not a constant that can be analyzed and reduced as easily as other facial aspects, such as bone structure and skin texture.

Machine learning typically fails here because, beyond the oft-criticized Facial Action Coding System (FACS), there is not enough empirical *terra firma* with which to curate and craft the data on which machine learning models might be trained to manipulate expressions.

Perhaps the most famous example of the possible dissonance between what a person is feeling and what others *perceive* them to be feeling is the poorly-named syndrome which we must here abbreviate as RBF, where individuals with nothing particular on their mind are perceived to be experiencing negative or 'aggressive' emotions, based on their facial expressions.

Likewise, the existing semantic logic of facial affect interpretation is currently still quite crude:

Three of the various possible effects – comic, intellectual and sexual – of the 'male raised eyebrow'; and even those are not definitive interpretations.

In the examples above, the *'male raised eyebrow'* creates three different impressions across three different examples, because the nearest psychological simplifications we have for this particular facial muscle movement are *'surprise'*, *'curiosity'* and *'disapprobation'*.

Therefore any machine system attempting to change the expression of an individual will need to understand the *specific* facial 'dictionary' that the person in question has developed over the course of their life.

Such a scenario is in opposition to the methodology of iterative machine learning systems, which are expecting to sift common and *consistent* features from the high volumes of data that they are trained on.

As people, we all have such idiosyncratic emotion>expression mappings that practically every example in a generically-labeled 'facial emotion' training database (beyond a broad beaming smile – and even that can be ambiguous) might qualify as an outlier, and defy *reliable* categorization; and you cannot train an effective AI system on outliers.

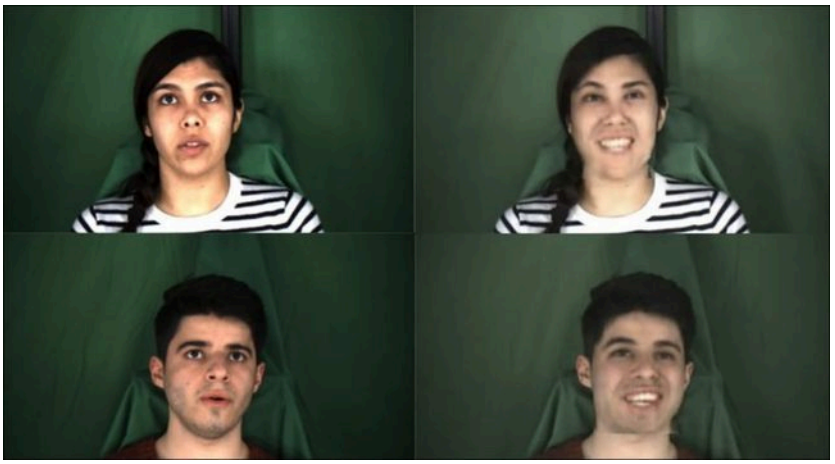
Changing Expressions

The ability to edit facial expressions is clearly in the roadmap for VFX adoption of generative workflows and methodologies. At the moment, regarding cinematic facial synthesis, the industry is placed roughly where Industrial Light and Magic was with motion-control work in the mid-1970s – determined to solve the biggest and most outstanding problems (such as tough angles), with the expectation that the standards for the (emerging) technology will rise over the years, along with the demands made on it by directors and other creative stakeholders.

At the moment, Metaphysic is breaking new ground in using neural faces to alter the ages of some of Hollywood's biggest stars, at full-screen resolution; but somewhere down the line, directors and actors alike will become accustomed to this functionality, and will begin to wonder if they can actually *fix* minor facets of an actor's performance *post facto*, such as 'dialing' an emotion up or down, or performing other affect-based transformations that are mere conjecture at the current state of the art.

Likewise, emotion recognition systems, whatever the motivation for their creation, would be greatly aided by a more granular and applicable understanding of how we facially express what we are feeling, and by the ability to quickly develop personalized and AI-based 'Rosetta stones' which accurately map the feelings>face pairings of a particular individual, instead of applying anime-style, generalized interpretations to a facial expression.

In the absence of this, and besides the shortcomings of Adobe Photoshop, most of the recent crop of expression generators, largely based on GANs, perform just as poorly, for the reasons stated. For instance, the 2021 project Wav2Lip-Emotion attempts to impose expression changes on video, but with questionable results:



Wav2Lip Emotion offers facial expression editing for video, but with artefacts and questionable accuracy. Source: https://raw.githubusercontent.com/IanMagnusson/Wav2Lip-Emotion/main/literature/Wav2Lip-Emotion_updated.pdf

Likewise the 2022 project *Neural Emotion Director* (NED) attempts to create live expression-editing facilities:

CLICK TO PLAY VIDEO ON SITE

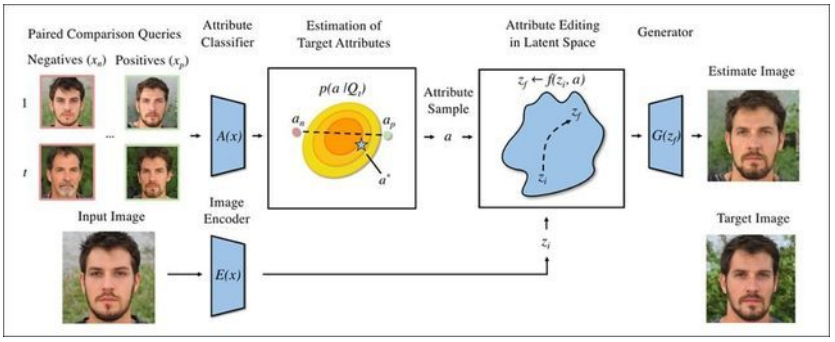
Disney Research has taken an intense interest in neural facial technologies in recent years, and its 2015 *FaceDirector* framework is probably the most successful attempt yet to develop an effective expression-editing method.

https://www.youtube.com/watch?v=o-nJpaCXL0k

However, this is not the 'purest' of approaches, and relies on the use of facial landmarks to normalize pixel-based warps, which are intended to define the change of emotion in the face. This is a laborious approach that adds little to the oncoming struggle to create individualized and meaningful expression-altering system through more modern methods such as GAN-inversion and straightforward image synthesis, with minimal reliance on CGI-based interfaces and other complex interstitial systems.

PrefGen

This topic has come to mind lately because of a recent paper released by the Georgia Institute of Technology, which offers a system that addresses some of the shortcomings of expression recognition, and is capable of editing expressions (among other facial aspects) based on real or simulated human feedback, rather than by nudging pixels around until they conform to some generic notion of 'happy', 'sad', etc.
Titled *PrefGen*, the new project can generate images based on human interpretation of whether an image of a face is happy, sad, angry, or any other troublesome or ambiguous emotion – because, in the end, despite cultural memes such as RBF, humans still have more innate and unschooled talent than machines at deciphering the language of the face.



The workflow for PrefGen. Source: https://arxiv.org/pdf/2304.00185.pdf

PrefGen works by presenting comparison queries to an *oracle* – a module within the framework which is trained to compare the pairs and select the most apposite one.

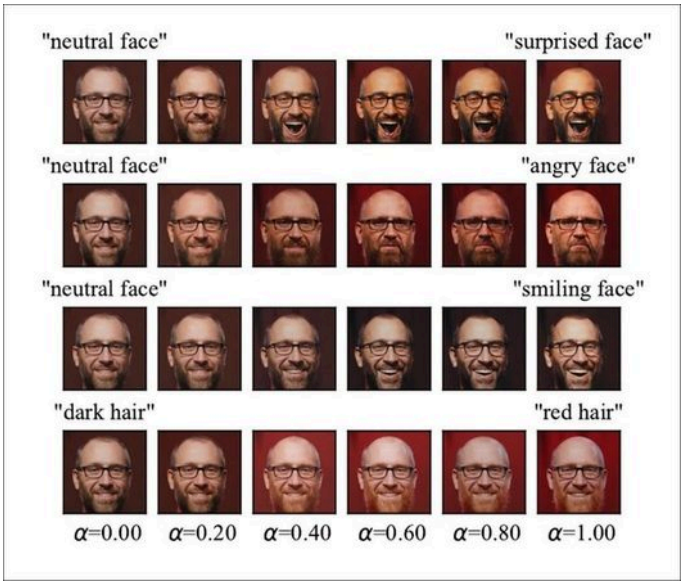
Though this may sound like just another automated loss function (or perhaps an LPIPS-style metric), the difference is that the choices can either be made by real people in real time, or (a more likely scenario), using crowd-sourced data that already features apposite human judgements which relate to the images.

The authors explain:

'[Many] attributes, like how "angry" a human face looks, are difficult for a user to precisely quantify. However, a user would be able to reliably say which of two faces seems "angrier". Following this premise, we develop the PrefGen system, which allows users to control the relative attributes of generated images by presenting them with simple paired comparison queries of the form "do you prefer image a or image b?"
'Using information from a sequence of query responses, we can estimate user preferences over a set of image attributes and perform preference-guided image editing and generation.'

The explanation further illustrates how arcane the pursuit of expression recognition is, since the authors suggest, by implication, that our best judgement in this regard may be essentially *subconscious* – perhaps developed as a survival mechanism, initially developed to assess possible threats or potential mates.

The system of consulting the oracle in PrefGen involves iterating through many pairs of faces, until an optimal solution is reached, and the chosen generated face is as close to the ground truth (known, real-world data) as it is likely to get.



Examples of modeling relative attributes (see below) in PrefGen using CLIP.

PrefGen gradually builds up an internal knowledge of user preferences regarding facial pairs until it arguably has something close to an authentic human opinion about the nature of the faces presented to it.

This results in a *face-specific* system of judgement, rather than an oracle that can make equally effective judgements about *multiple* identities. The estimation of target attributes carried out throughout PrefGen's operation will need to be repeated for different identities, so that a bespoke editing system can be created in each case, based on human judgements about that what the subject's facial expressions signify.

This is a relatively labor-intensive process, even when the data pairings can be obtained from external sources (which limits the oracle's capabilities to identities that are associated with that public data – 'transferability', the ultimate aim of AI-based training and automation, is minimally possible in this scenario, for the above-stated reasons).

Approach

Though, for the purposes of the paper, PrefGen uses a StyleGan2 architecture, the authors state that the core principles of the framework could be applied to other types of generative system.

The system trains supervised mappings from facial attributes such as age and pose (facial expression and facial direction, etc.), in accordance with the methodology of the 2021 Amazon initiative GAN-Control.

Amazon's GAN-Control framework allows control over the disposition of facial images in the latent space of a GAN. Source: https://arxiv.org/pdf/2101.02477.pdf

The system takes a couple of approaches to face editing, one of which leverages the hugely popular OpenAI CLIP model, which establishes relationships between pictures and text, so that text can be used to control and edit output. Effectively the user-based judgements add a third dimension to the existing growth of multimodal systems such as CLIP, by adding an 'unconscious judgement' to the image/text pairings that typically comprise such systems.

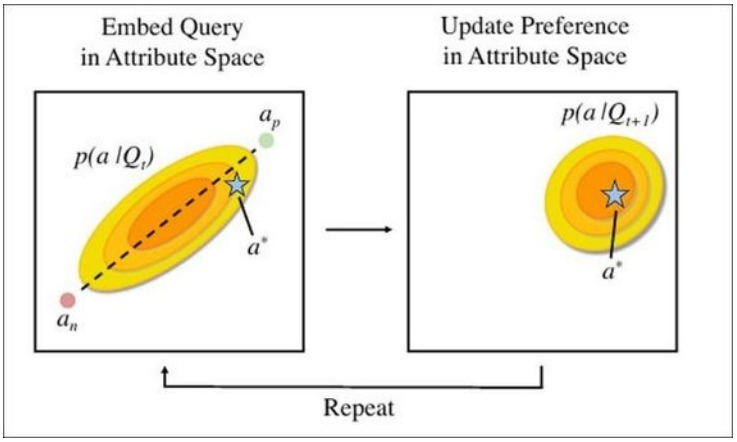
PrefGen is predicated on the notion of *relative attributes*, a concept first presented, in this context, in a 2011 paper from the Toyota Technological Institute Chicago and the University of Texas at Austin.

From the 2011 paper about Relative Attributes, some examples that demonstrate how binary labeling and classification can fail in ambiguous cases. For instance, in e in the image above, to what extent does the image express the idea of the concept 'manmade'? Source: https://www.cs.utexas.edu/~grauman/papers/ParikhGrauman_ICCV2011_relative.pdf

A machine learning purist may consider the ambiguity of relative attributes as simply a stubborn problem that later advances will eventually solve; but the new paper is predicated on the assertion that binary classification is too crude a tool for many 'edge' cases, or in domains such as facial expression, where constants are lacking, and where only prior knowledge about the development of an individual could really help to create systematic emotion/face mappings – data that is unlikely to be available, except through precarious 'guesswork'.

Therefore the authors are essentially conceding that there are domains which *may never yield* to broad systematic analysis, but which will always require per-case processing – albeit that the processing methodology will in itself be consistent and 'templated'

In any case, once the preferences from paired image comparisons are in, the distribution representing a user preference is turned into something more tangible – updated preferences in attribute space, which can be applied in a generative or editing system.



Rationalizing the irrational, as user preferences are turned into vector data which can be used in a generative or editing system.

Though the new paper is quite exhaustive on the conceptual methodology, it does not outline how time-consuming the act of manual user choice would be, or how many times it would have to be gone through to generate a per-case algorithm for editing a particular face. The authors refer to prior work from Cornell which explores the exploitation of web-log data to infer user preferences, and suggest that such a method could be used to automate an oracle, rather than requiring direct user input.
They state:

'[It] is possible to synthesize user preference information from abundantly available sources and use generative models to generate content that matches user preferences.'

However, this would clearly only work in cases where the face in question was adequately well-known that a high enough volume of data would be available from which to derive these preferences – i.e., celebrities and influencers. For the 'obscure face', it seems that manual intervention would be the only rational way forward. What works for Tom Cruise is unlikely to be as applicable to 'John Doe'.

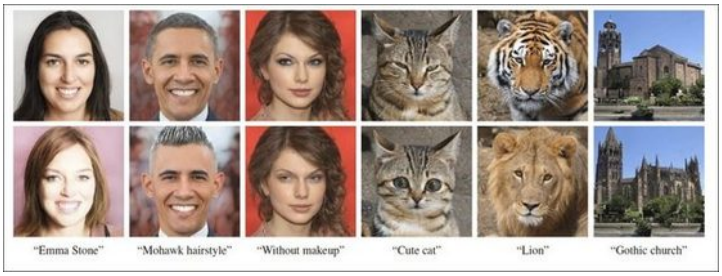
The ultimate goal of the paired comparisons used in PrefGen is to arrive at a common simplifying assumption for the identity in question, called an *ideal point*.

To estimate this point, the authors draw heavily on 2019 work from the School of Electrical and Computer Engineering at Georgia Institute of Technology, which offered a metric for selecting queries from a dataset while accounting for high 'noise' in the responses (i.e., non-pertinent responses). Using this method, PrefGen selects optimal queries from a continuous stream of data, instead of the fixed sets used in the 2019 paper.

PrefGen also draws notably from 2019 research from the US Air Force Research Laboratory at Rome, which uses Long Short Term Memory (LSTM) to predict a latent vector from the image-pair responses. However, the authors denounce this method as inefficient, using it as a baseline for the project's experiments.

Tests

PrefGen uses two approaches for image editing – one based on Amazon's aforementioned 2021 GAN-Control paper; and one based on StyleCLIP, an Israeli/Adobe collaboration from the same year.

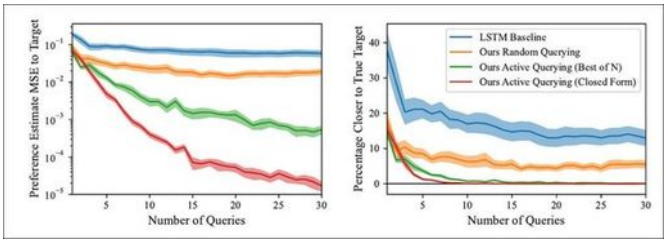


StyleCLIP uses text/image pairs as a method of altering faces in a StyleGAN framework. Source: https://github.com/orpatashnik/StyleCLIP

The authors have used StyleCLIP's approach to interpolate between a neutral and target text-prompt, such as *'person with neutral expression'* and *'person with angry expression'*.

The researchers conducted a raft of tests for PrefGen. The testing criteria adopted by GAN-Control was used to construct supervised facial mappings covering yaw, pitch and roll of the face, as well as perceived age, using the FFHQ Human Attributes Dataset. Twenty trials were conducted across the various methods, each with 30 queries.

For the comparison to a baseline method, they used the Air Force Research paper mentioned above as a baseline.



Results for the LSTM (US Air Force) baseline vs. PrefGen.

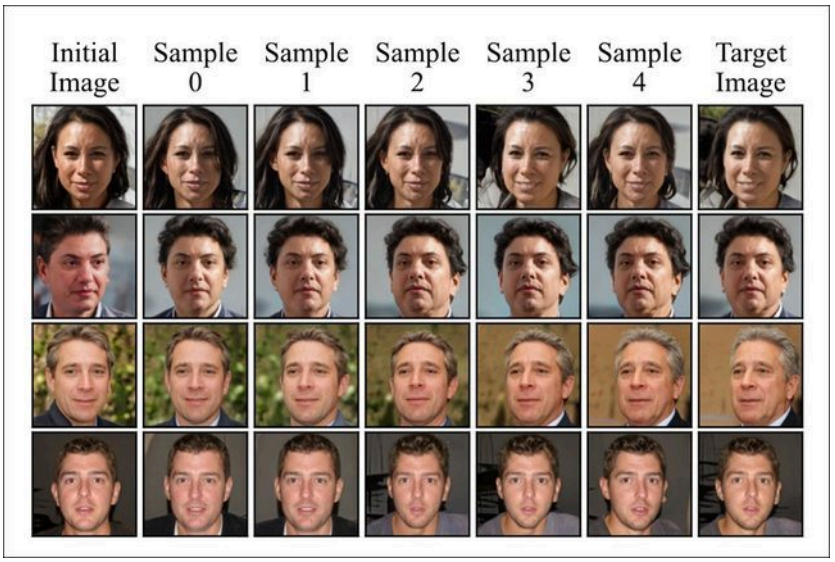
Of this part of the tests, the authors state:

'In our approach, we separate preference estimation and mapping to the generative model latent space. For the sake of comparison, we extend the LSTM approach to estimate a low dimensional attribute vector that satisfies a given set of constraints […] [Our] method with both random and active querying can estimate the target attributes with higher precision and efficiency.'

For a quantitative evaluation, the authors used three metrics: Mean Squared Error (MSE); the extent to which the attributes obtained were closer to the target attributes than the authors had anticipated; and the percentage of paired comparison constraints which veered towards the positive end of predictions. These results are also visualized in the image above, and the authors assert that PrefGen obtains a (good) low MSE score, satisfying a high percentage of constraints.

For qualitative tests, the researchers conducted two trial runs for generating a given preference estimate:

CONTINUED ON NEXT PAGE

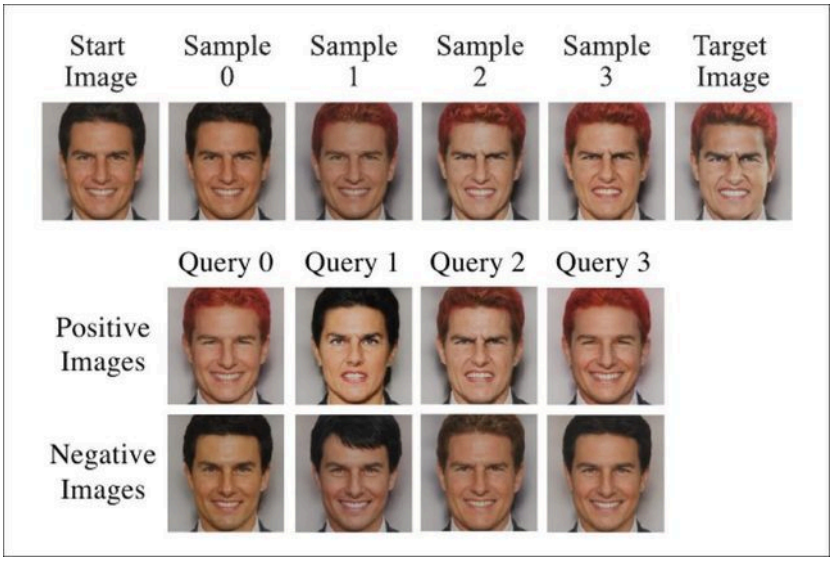


The images generated here match oracle preferences in PrefGen, gradually approaching the qualities of the target images as the system iterates.

The authors comment:

'Qualitatively, the attributes of the preference samples converge to the attributes of the target images…Qualitatively, it is clear that the preference estimate is converging on the target attributes using the information provided by each paired comparison query.'

The authors also experiment with face-editing, using the Encoder4Editing framework, successfully editing the facial vectors obtained by the oracle's judgement on the face pairs.



Here the oracle is estimating the user's preference for the textual attributes 'red hair' and 'anger', using a sequence of pairwise comparisons. The lower two rows show queries presented to the oracle, while the top row shows the generated estimate of the oracle's inferred preference after answering each query. Effectively, an overall 'opinion' builds up in this way.

For further results, in an extensive round of testing, please see the paper.

Conclusion

The authors of PrefGen are boldly addressing an emerging challenge in facial synthesis, based on the assumption that neither the growth of hyperscale datasets nor the possibility of code-based innovation can solve the problem of personalized emotion editing (among other slightly less challenging tasks, such as rotating a head slightly).

This assumption is a cultural one, and may be considered an open question – are we universally 'readable', in terms of facial expression significance? And if AI were to aggressively tackle the problem (which is, for the moment, largely on the back-burner), would it unearth universal constants that could accurately equate a facial expression from any individual with the emotion underpinning it?

Though it's a fascinating anthropological question, the immediate way forward for low-shot accurate neural expression editing would seem to be using something at least similar to the PrefGen approach, where human instinct is used as a proxy for what is currently a very nascent field of study in psychology itself.

Even for well-documented individuals such as Tom Cruise, or other celebrities, there simply exists no truly reliable applicable framework that can perform this mapping. To an extent, creating systems such as PrefGen may be putting the cart before the horse; we need much deeper insights into the semantics and logic of facial affect before we can begin to quantify such information accurately into generative systems.