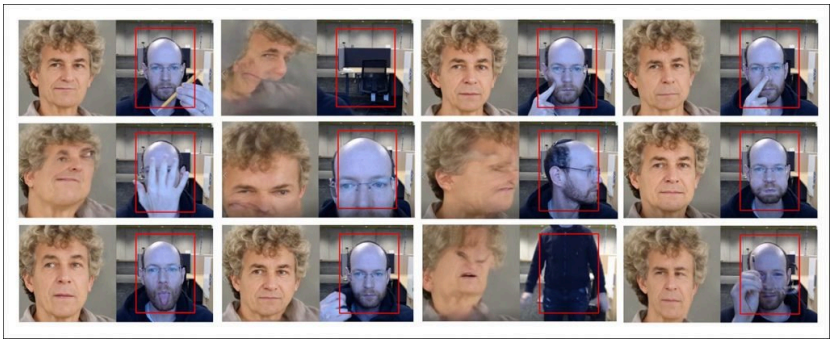
Are We Heading For 'Deepfake CAPTCHA' Challenges in Video and Voice Calls?

Originally published at https://metaphysic.ai/are-we-heading-for-deepfake-captcha-challenges-in-video-and-voice-calls/

August 23, 2022

A new paper from Israel proposes the institution and ongoing development of a 'deepfake CAPTCHA' protocol, to challenge audio or video callers that may be using deepfake technologies to attempt to deceive or defraud.
The initial challenges suggested by the work include a test that mimics our recent proposal to use the inaccuracy of deepfaked profile views as a metric of authenticity for a potentially deceptive live communication across platforms such as Zoom, Skype, and FaceTime.
The paper – which is titled *DF-Captcha: A Deepfake Captcha for Preventing Fake Calls* and comes from Dr. Yisroel Mirsky at Ben-Gurion University's Offensive AI Research Lab – goes further, proposing a slate of challenges that are in many ways similar to a traffic cop's 'drunk tests', characterized by the author in terms of a novel twist on a traditional Turing test.



Some of the proposed challenges to determine whether a caller is being impersonated. The new paper also includes tests to probe the limits of voice-based deepfakes, the sole source of data in an audio-only call. The paper explains that it outlines concepts filed in a patent application submitted on 1st January 2022. Source: https://arxiv.org/pdf/2208.08524.pdf

Assume the Position

The new work envisions automatic systems built into communications channels, which may – on request or automatically – ask to test one of the communicants for the possibility of deepfake content.
For video, the initial round of tests proposed, which could be requested in their entirety or used as a resource from which to choose a smaller number of specific tests, include *drop object*, *bounce object*, *stroke hair*, *crease shirt*, *interact with background scenery*, *spill water*, *vibrate lips*, *show a* (requested) *object*, *remove glasses*, *wave hand quickly in front of face*, and *perform tongue motions*.
Some of these seem to be aimed at potential full-synthesis systems that exceed the scope of autoencoder-based deepfakes, which in themselves only replace the inner contents of a subject's face.
Others, such as *stroke hair*, would indeed prove a challenge for the growing practice of full-head deepfakes (see embedded video below), which incorporate the entire cranial area, including the subject's ears.

https://www.youtube.com/watch?v=eWATIYiQZeo

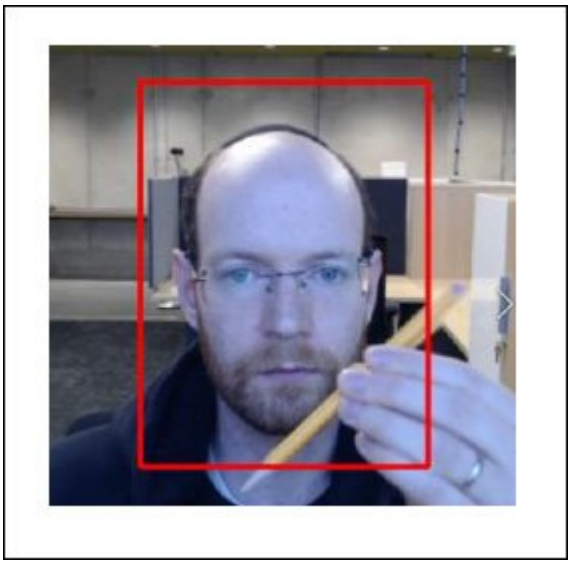
Hair simulation is currently one of the sharpest challenges in identity simulation, and the current state of the art suggests that we are many years away from solutions that would run effectively in well-specced VFX house systems, never mind on domestic GPUs.
However, in most cases, the attacker's hair is likely to be real, as is every part of their video feed with the exception of the inner facial area, and, to a more challenging extent, the outer lineaments of the face, such as cheek and jaw edges, and profile details.
Let's take a look at the rationale behind some of the initial proposed challenges.

Logic of Challenges

As the paper points out, the roster of potential challenges can be added to over time. This is because, in the deepfake video-call scenario, the attacker is very much on the back foot, needing to train a model to anticipate an ever-wider range of contingencies, and in many cases to develop synthetic data to accomplish this.

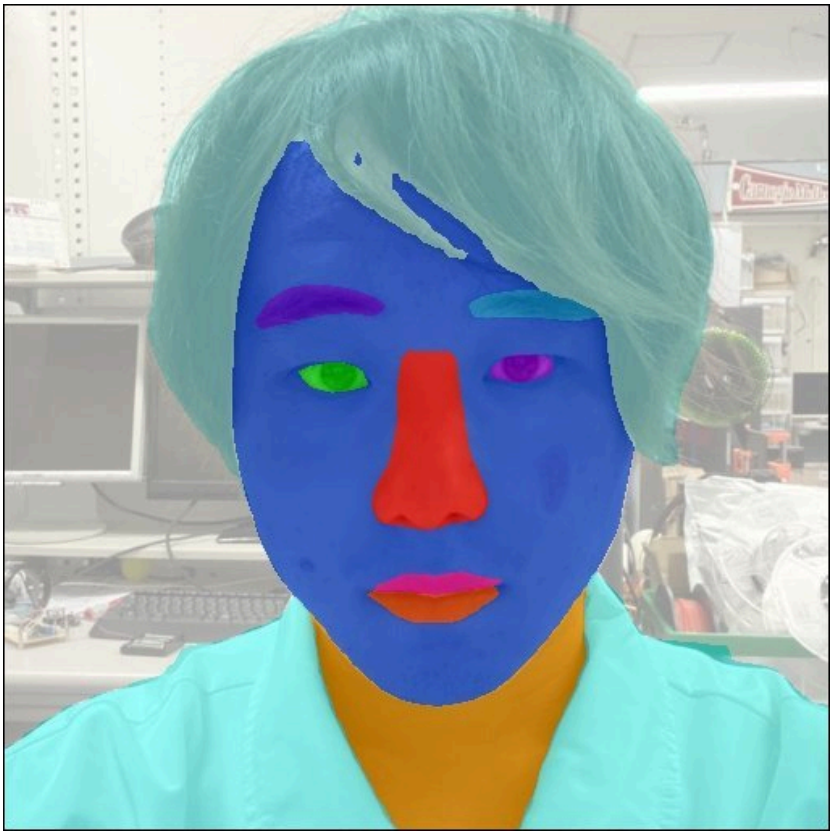
The Pencil Test

In the case of pick up an unknown object, we can see in a detail from the paper's examples that the subject is being asked to hold the object in front of his face.



Unless the attacker has specifically trained the model to accommodate a wide variety of facial occlusion objects, this challenge is going to reveal matting artifacts immediately. The proposed detection routine should easily be able to detect such artifacts, as, indeed, should the other communicant in the call.

Autoencoder-based systems such as FaceSwap and Machine Video Editor make use of the BiseNet semantic segmentation network, which, depending on the training data used, can allow for a wide range of facial impediments of this type.



The BiseNet segmentation network can individuate parts of the face, and can also help to train a machine learning model to 'cut out' objects that may occlude the face. Source: https://www.researchgate.net/figure/An-example-of-face-segmentation-using-BiseNet_fig1_359685832

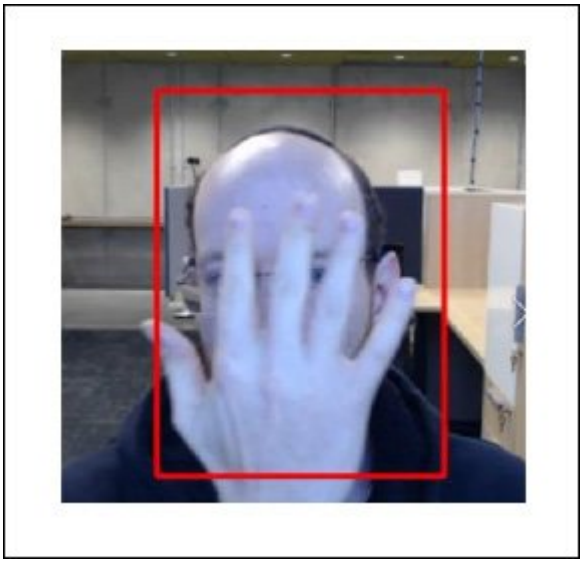
However, neither BiseNet nor similar occlusion-handling networks can account for every possible object that the other communicant might request, and is going to do best with 'obvious' facial occlusions such as glasses and microphones. Further, including a truly comprehensive range of objects will not only increase training time, but decrease available attention and model accuracy for other aspects of facial recreation.

Further, if you wanted to ensure that your deepfake attack model could stand up to matting out a pencil held in front of the face, you're either going to have to train BiseNet (or whatever segmentation network you're using) with specific examples of that kind of occlusion, and you'll likely need to train 'adulterated' face images (i.e. you'll need to create many training images in which an artificial hand/pencil combo is obscuring part of the face).

And that's just pencils; and that's if the *synthetic* data (no real data will be available for such a marginal case) can adequately train for the 'real thing', which is not guaranteed.

Talk to the Hand

One object pretty much guaranteed to be available is the attacker's hand. Asking him or her to wave it in front of their face will further stress the trained deepfake model, which either will not have had access to such images during training, or will have only had access to synthetic data, with all the accompanying disadvantages of the 'pencil test'.



In our recent article on profile-based deepfake detection, we took a look at the reasons why deepfake models cannot easily, accurately or quickly reproduce this interaction:

No Tongues!

In common with many facial synthesis systems currently under development, autoencoder-based deepfakes offer scant account for the *inner content* of the mouth. In most cases, only the teeth will ever get resolved during training, and perhaps a more-or-less generic red representation of an area of color intended to represent the tongue.
This means that forcing your potential attacker to perform tongue-based manipulations is going to present a real challenge to the deepfake model, which has been exposed to practically zero data of this type.

It's not that a deepfake model couldn't learn to recreate tongues; it's just that the data is not only understandably scarce but also extremely difficult to synthesize. As our recent contributor, expert deepfaker Deep Homage has commented in the Machine Video Editor Discord, *'For typical training, there won't be enough tongue examples for the model to learn'*.

I'm Really Touched

The next example included in the select images from the paper is *touch face*.

No deepfake attacker is likely to attempt a video-call deception with a model that can't handle at least finger occlusion. But can it handle reproducing the distorting effect of pressure on the face, across a range of possible lighting conditions, and with movement possibly included?
Of all the examples so far, this is probably the most difficult to 'retrofit' onto training images of faces, since it not only would involve the elaborate superimposition of hands in all sorts of configurations into the training images, but would also require the addition of hyper-realistic geometric facial distortions and their associated interior shadows – a task that would challenge even the post-processing departments of visual effects houses.

Glasses Off

Models created for an autoencoder deepfake frameworks such as DeepFaceLive are often trained to occlude face-worn glasses quite well, though it does take a notable amount of training data to get a really good result.
However, the act of *removing* the glasses occupies only a few scant frames of any potential source video (making the necessary training material very hard to find); requires complex matting, including accounting for motion blur; and may omit obvious specular effects relating to the target lighting conditions, which may not have been trained into the model.

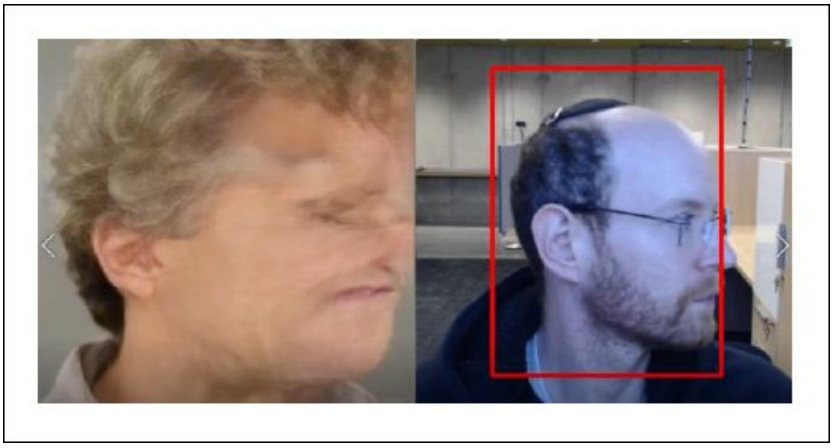
Hello, Duck Face

Facial contortions, either with external pressure (as in the *touch face* example above), or though straightforward gurning, are inevitably going to be out-of-distribution (OOD) data as far as the trained deepfake model is concerned.

Though it's possible that some subjects could have provided limited types of unusual facial distortion data on social networks, the training data will not be able to account for all the possible variations of facial distortion that a detection system might require of a potential deepfake attacker.

Eyes Right

Here, the paper comes to the same conclusion that we did in our recent feature on profile-based deepfakes: that for various reasons, primarily the lack of data, deepfakes are very bad at recreating oblique profile views. The new paper includes this among its slate of initial recommendations for a deepfake Turing test.



Our own tests, conducted with the help of tech exponent Bob Doyle, confirmed that the difference between an 80° and a 90° turn to profile is the difference between a passable deepfake and a hot mess of hallucinated pixels, in the case of most deepfake models:

We have argued that this approach to live deepfake detection is particularly effective due to the scarcity of profile data available for any potential deepfake victims that have *not* carved out a career in movies and TV.

Audio Deepfake Attacks

The new paper radically reframes the atmosphere of terror inspired by recent scandals around 'live' deepfake scenarios, suggesting that a framework of ever-growing challenges will put potential Zoom-based fraudsters into a defensive stance. The paper states:
'What makes this technique powerful is that the challenge is easy for a human to perform but extremely hard for a deepfake model to generate. For example, to expose facial reenactment, the challenge might be to have the caller move his/her head to an oblique angle, press on the nose, or simply turn around.'
Some of the most headline-grabbing attacks have either made use of audio deepfake content in videoconferencing, or have entirely depended on audio. For instance, the 2020 incident – which came to light the following year – in which a Hong Kong bank manager was persuaded to transfer $35 million USD for a supposed company acquisition, but which transpired to have been an elaborate fraud enabled through deepfake voice-cloning technology.

To demonstrate the relative ease of creating such audio simulations, the paper links to the author's own simulation of a potential voice attacker, created by the method outlined in a prior research paper from Google.

As with video-based deepfakes the new paper suggests a range of possible tests for voice-based fraud, including vocal events that are very unlikely to have been trained into the model. Such tests could be applied either as a secondary line of defense in a potential deepfake video/voice fraud attempt, or in a voice-only context, such as a voice-based Skype or Zoom call – or even a regular telephone call.

Some of the initial tests proposed by the paper include *mimic phrase*, *hum tune*, *sing part of song*, *repeat accent* (which may be challenging for the less gifted mimics among us), *change tone or speed* (of voice), *clear throat* – or *whistle*.

It should probably be considered that, except in rare cases, the potential victim may not be familiar with certain possible perturbations of the attacker's voice, such as throat-clearing, or the quality or tone of their singing voice, and that generic substitution in the training set could arguably block these detection vectors.

However, vocal contortions that leave the core voice intact, such as accent attempts, the pronunciation of unusual phrases (for instance, foreign language phrases that are unlikely to have been included in the model's training data), or changes in tone, seem likely to remain effective in any putative automated detection system.

Focused Attention

The paper's author observes that a CAPTCHA-style system of this kind has a notable advantage over traditional architectures, in that it is not obliged to dumbly monitor an entire stream of content for potential anomalies or 'tells'. Rather, it can prompt a specific test from the challenged communicant, and force the test material to take place within a time-frame and under circumstances that the system, rather than the challenged user, controls.

The new paper also illustrates the extent to which the security research community has until now concentrated on artifacts that may reveal the deepfake process, such as mesoscopic properties, spatial-domain notch filtering , sequential patches, frame-level features, and transferable distribution characteristics, among many other frame-based and temporal approaches.

The locus of effort in deepfake detection has until very recently concentrated on the potential societal impact of fully-rendered deepfake videos, which can enact their deceptions on their own terms, and which are relying on the (arguably diminishing) credulity of a general public that still believes the camera never lies.

However, if you can make a potential deepfaked caller do whatever you want for at least a few seconds per test, the detection proposition begins to significantly favor the supposed victim rather than their attacker.

There are, in effect, no limits to the hoops that such an approach can make a caller jump through in order to authenticate them, while there are very definite limits both on the availability of training data that could encompass such a wide range of bizarre requests, and on the ability of a performant model to integrate and execute them flawlessly.

The only major problem to enacting deepfake CAPTCHA systems over popular communications channels is that the current crop of 'concerning incidents' may yet be too infrequent and marginal to force the change. Sadly, it may be that we need a bigger headline.