Identity in the age of deepfakes

or How we learned to stop worrying and love deepfakes

Whether we like it or not, deepfakes are now part of our daily digital lives. What used to require tons of compute and very advanced skills has now become a commodity that any programmer can leverage with a laptop and an off-the-shelf GPU.

The last remaining barrier was data. The presumption was that even the best model needed specific data to generate data useful for an identity theft (a face biometrics or a government ID).

This assumption is unfortunately defeated by empirical evidence: with the right amount of prompting, both open-source and proprietary models are now able to generate highly realistic identity samples. Below are two examples. These government IDs never existed and would arguably fool a human analyst. Is the identity industry doomed?

Documents generated with ChatGPT

We have good news for you: the answer is an overwhelming no. At Onfido, we saw this threat coming years ago and acted decisively against it, for the benefit of our customers. We could do it because we are equipped with a unique combination of machine learning expertise and an in-house Fraud Lab, both totally 15 years of industrial experience in the fight against fraud.

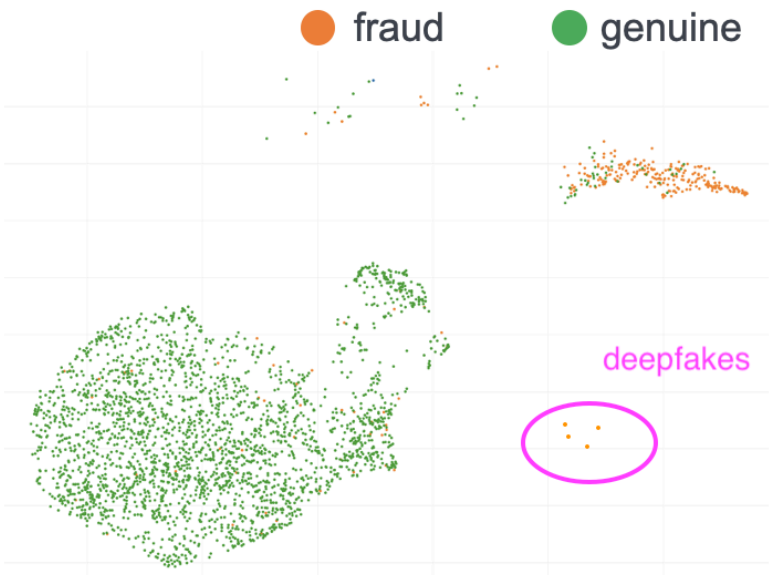
Here is the first deepfake we saw in production. We call this one "deepfake patient 0". We spotted it in 2021 and built our flagship Motion product anticipating such deepfakes at scale in production. The future proved us right.

Let us explain why these deepfakes are far from being as dangerous as they are (at least if you use us as an identity provider) and how we stay ahead of the curve.

Early detection is key: Fraudsters use a trial-and-error approach. They first try a new approach at a small scale before going gung ho with it. Our continuous monitoring of production gives us the ability to catch these frauds early on. One would be tempted to think that this is a needle-in-the-haystack problem, which turns out to be true. However, we have built machine-learning based, proprietary methods to detect these deepfakes in the massive amount of genuine data.

Deepfakes are out-of-distribution for machine learning models: As realistic as the GenAI-generated samples may be, they do not present the visual defects due to the

physical process of capture through a cell phone camera. In other words: they look too perfect. Our models are trained on large amounts of real data and therefore consider these samples to be out-of-distribution. To fool our models, fraudsters would need to not only generate highly-realistic samples, but also simulate the physical process of capture.



Machine learning model project samples into an "embedding" space. Samples that have not been captured by an actual device land far away from the genuine space, making them suspicious to the model.

Deepfakes leave traces behind them that are invisible to the human eye but visible to the machine. Those can be very small artefacts at the pixel level as well as unusual noise patterns in the image. A well-maintained supervised learning engine trained on millions of samples spots these trails and catches the majority of deepfakes. This approach is heavy, though: it requires a lot of data, constant retraining and continuous generation of new fraud samples. It also goes against valid privacy concerns. We will therefore keep doing this for the foreseeable future but are also working on the long game – secret sauce in the making!

Deepfakes require an injection attack: Fraudsters who generate deepfakes need to "inject" these samples into the device used for identity verification (usually, a

smartphone). This requires the fraudster to bypass security features of cell phones, which become harder and harder as manufacturers raise the bar. We provide an additional layer of protection using a proprietary, in-house device intelligence that we improve continuously. Alternatively, fraudsters may attempt to use a genuine phone to capture a photo or video of the deepfake. We call this a "video-of-video" and have developed specific machine learning models against them. Much like deepfakes, screen captures leave small trails behind them that our models catch robustly.

Fraudsters cannot hill-climb with deepfakes: Hill-climbing is a technique used by fraudsters to continuously improve their attack based on rejections from the system. Hill climbing is useful when fraudsters can build a mental model of what is wrong with their failed attempt. Counter-intuitively, generating highly realistic deepfakes makes hill-climbing very hard. If a highly realistic deepfake was rejected, then what goes? We observe the consequence in production: fraudsters need to perform a large number of repeat attempts with many variants of their models. This is where our Repeat Attempt product kicks in 🙂

Design Secret sauce: We design our Identity products from the ground up and entirely in-house. This opens the door to a significant amount of secret sauce, building the UI/UX, the SDK-based capture process and the machine learning models together in one go. This approach allows us to reach an unmatched combination of low friction and fraud protection. These designs represent a significant value proposition for our customers.

In conclusion, while deepfakes make the headlines in the press and represent significant work for us, we provide strong protection against them thanks to:

- Early detection and continuous monitoring
- Dedicated machine learning models
- SDK-level injection attack protection
- Proprietary and constantly evolving product designs

Prediction for the future

Fraud is becoming a scaled, organized business. Just do the math: if a bitcoin company offers a $50 voucher for an account opening, opening 10,000 fake accounts brings you $50K in a few hours. Not a bad day.

Targeted account takeovers are also becoming attractive. Careless people still use weak passwords. More careful people use a proper 2FA. The majority of people won't go through the hassle of using hardware-based security (think yubikey). This leaves the door wide open for well-prepared software-based attacks.

Deepfakes on government IDs and biometrics will become even more prevalent. As deepfake generation has become a commodity, other fraud vectors are just harder to produce at scale. Our multi-layered, constantly evolving protection will continue to protect our customers and their end users.

Asymptotically, the fight is moving on-device. Deepfakes require an injection attack. Ensuring that the user is using a genuine device will be key. Entrust's extensive knowledge of device security comes as a strong complement to Onfido's machine-learning based technology.

Data moats remain more useful than ever. As we reach a very high level of performance, just *measuring* performance numbers requires very large volumes of data. In addition, fraudsters' early attempts are often un-noticeable at small scale. The data moat we have built over the past decade gives us a unique edge against deepfakes.

In short, while deepfakes can be worrying from the public standpoint, the landscape in industry is very different: we provide strong protection against them. We work hard for our customers so they don't have to.

www.entrust.com