Yoti's response to the threat of generative AI

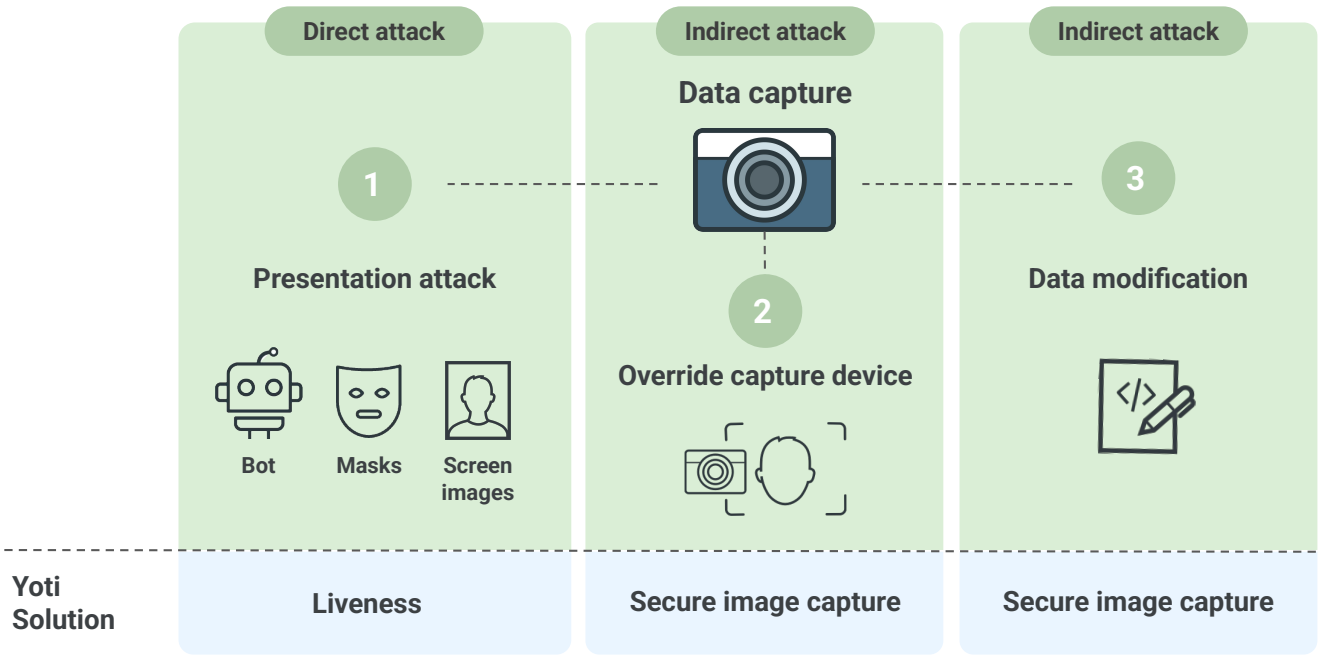
January 2026

Yoti's response to the emerging threat of generative AI

Our strategy for detecting the threat of generative AI covers two vectors of attack. This approach focuses on early detection; i.e. the point of the attack a bad actor is attempting to enter a generative AI image or video into the verification process.

Bad actors will try to spoof the image input into the verification process in two ways. The two vectors of attack are presentation attacks (direct attacks) and injection attacks (indirect attacks).

Data capture attack threats



Presentation attacks

- **Relatively mature and well understood issue across the verification space**
- **Presenting an alternative image, video or mask (for example) to a live camera**
- **A more naive / more obvious attempt to spoof**

A presentation attack is an attempt to spoof a verification by pretending to be someone else to the device camera. This could be using a screen image, mask or generative AI.

The purpose of a liveness check is to make sure the person you are verifying is a real person. Attempts to spoof a verification in this way are often termed as a ‘direct’ attack.

Yoti’s proprietary passive liveness detection technology, MyFace, meets iBeta Level 3 liveness standard - the highest liveness standard for presentation attack detection. MyFace is a passive solution and only requires a selfie image to complete the liveness check, creating a smooth, inclusive and accessible user experience.

We also offer businesses the option to use two other leading liveness suppliers that offer a fall back and / or complement to Yoti’s proprietary system. This can be included as part of the process as either a ‘double check’, a fallback or for where ‘high’ risk has been identified.

Read more about liveness and MyFace here.

Injection attacks

- **An increasing threat vector**
- **Bypassing the device camera to 'inject' an alternative image or video**
- **Basic tech understanding and investment to attempt an attack**
- **Recommend to be used as a complementary product to liveness for generative AI detection**

An injection attack is an 'indirect' attack and attempts to bypass liveness detection. It involves injecting an image or video designed to pass authentication, rather than the one captured on the live camera. It is a rapidly emerging threat to digital verification services. Using free software and some limited technical ability, a bad actor is able to override the image or video of the camera with pre-prepared images.

This can take two forms

- **Hardware attacks**:
  this method connects directly to the camera of a device, with the video played from another device. Effectively, this replicates a live video with a generative AI video injected at this point.

- **Software attacks**:
  this could be:
  - virtual cameras, that simulate a physical camera as if it were a real webcam, or;
  - exploiting breakpoints in the device or browser SDK code, to insert an alternative image or video not taken from the live camera.

Yoti have developed SICAP (**S**ecure **I**mage **CAP**ture), a patent pending solution that makes injection attacks considerably more difficult for imposters. It is a new way of adding security at the point an image is being taken for a liveness or facematch check. SICAP can detect both software and hardware attacks.

There are two parts to how SICAP works. As well as obfuscating the code at the point an image is taken, Yoti adds a cryptographic signature key. As such, a potential hacker needs to both reverse engineer the obfuscation and infer or guess the cryptographic signature key.

Yoti frequently changes the obfuscation and the signature key. This means that if the hacker were to reverse engineer the obfuscated code, by the time they have done so, the signature key will have changed, and vice-versa.

We also block by default virtual camera software such as ManyCam, VCam, vclsCam, FakeWebCam.

By using both liveness and SICAP products, you can be sure you have the latest technology available to combat attempts to spoof and or produce generative AI content.

Developments - SICAP 2.0

Our latest update for SICAP 2.0 is able to detect both hardware and software attacks.

We have improved our virtual camera detection with different methods of fake camera detection.

We have allowed organisations to opt for whether to receive the image or not, which is better for user privacy and data minimisation. For example, if an organisation requires an age check, there is no need to receive the image. If it's an identity verification, organisations will in most instances require the image for their customer files.

We have also now localised SICAP in a total of 40 languages.

Finally, we have improved analytics and support for integrators, including *client-side drop-off* analysis to help identify potential problems.

Risks

At present publicly available or understood generative AI models have been developed by researchers or benevolent commercial organisations with the goal of tricking the human eye - it is not being done with malicious intent.

However, combatting generative AI has now become an arms race. We can be sure states, state sponsored actors or sophisticated criminal organisations have developed or are working on generative AI tech. No-one really understands how generative AI will develop.

Once evading detection is added to a generative AI model as an objective during training, things will very quickly get a lot more difficult, particularly with images.

Methods to detect generative AI videos still have some longevity, as they have the advantage of being able to detect inconsistencies in the temporal domain. It's currently very difficult to achieve temporal consistency when generating generative AI videos.

Alternative approaches

There are other organisations and methods of offering generative AI detection but we have not been able to verify their accuracy. It has also recently been proven at the CVPR 2023 4th anti spoofing workshop that the detectable elements of generative AI can be easily removed, deprecating the effectiveness of this detection technology very quickly (reference). To refer back to our strategy, the pace of generative AI innovation can quickly make any good / new generative AI detection quickly redundant.

Hence our strategy and approach of targeting the problem at source as described above.

To learn more about Yoti's approach to combating Generative AI please **get in touch**.