



On the threat of Generative AI

February 2024





Executive summary

As the threat of generative AI in identity and content integrity continues to build, Yoti has developed a comprehensive strategy focused on early detection by using tools to prevent AI-generated content or attacks at the point of source.

Yoti's strategy for detecting generative AI threats targets two attack vectors: presentation attacks (direct) and injection attacks (indirect), with a focus on early detection during the verification or authentication process. We will explain how requiring consent on media platforms for problematic images or notable individuals can also help mitigate unwanted or undesirable imagery or video.

The rate of development of generative AI presents a problem to not just ensuring a person is who they say they are, but also to content platforms who need to be sure that the content added by a user is genuine. This could be video sites, live streaming, dating profiles or social media platforms.

Given the potential risks and challenges in detecting generative AI, Yoti's strategy emphasises early detection at the source, addressing both direct and indirect attack vectors. The integration of Yoti's MyFace and SICAP, along with ongoing advancements, positions Yoti at the forefront of combating evolving threats in the generative AI landscape.

We aren't suggesting we have the whole solution, it would require a collaborative effort across different parties.

- **Platforms** - will need to implement further policies that would drive away or suppress bad actors whilst simultaneously increasing trust, with very little effect on friction
- **Video and image creation tools** - efforts are also underway by image and video production tools to add digital watermarks to content. This will help with authentic content, but still the issue remains for inauthentic content.
- **Creators** - buying in to new processes to ensure their intellectual property or indeed image is not compromised.

The growing challenge

While offering numerous benefits and advancements in various fields, generative AI also brings certain potential threats and challenges. Some notable concerns associated with generative AI include:

Deepfakes and Misinformation:

- Issue: Generative AI can be used to create highly convincing deepfake content, including images, videos and audio recordings.
- Threat: This raises concerns about the potential for misinformation, fake news and the difficulty in distinguishing between authentic and generated content.

Identity Theft and Fraud:

- Issue: Generative AI can be leveraged to create synthetic identities and realistic forged synthetic documents.
- Threat: This poses a risk to identity verification systems and may be exploited for fraudulent activities, including financial fraud and unauthorised access.

Privacy Concerns:

- Issue: The ability to generate realistic images and videos raises privacy concerns, especially when it comes to creating fake content that features individuals without their consent.
- Threat: Privacy violations may occur as people's likenesses are used in ways they did not authorise.

Security Risks in Authentication Systems:

- Issue: Generative AI can potentially be used to fool biometric authentication systems.
- Threat: This could compromise the security of systems relying on facial recognition, fingerprints or other biometric data for user authentication.

Cybersecurity Attacks:

- Issue: Generative AI can be used to create sophisticated phishing attacks, including fake emails, websites or messages.
- Threat: Cybersecurity threats can become more challenging to detect and mitigate as attackers leverage AI-generated content to deceive users.

Challenges in Content Moderation:

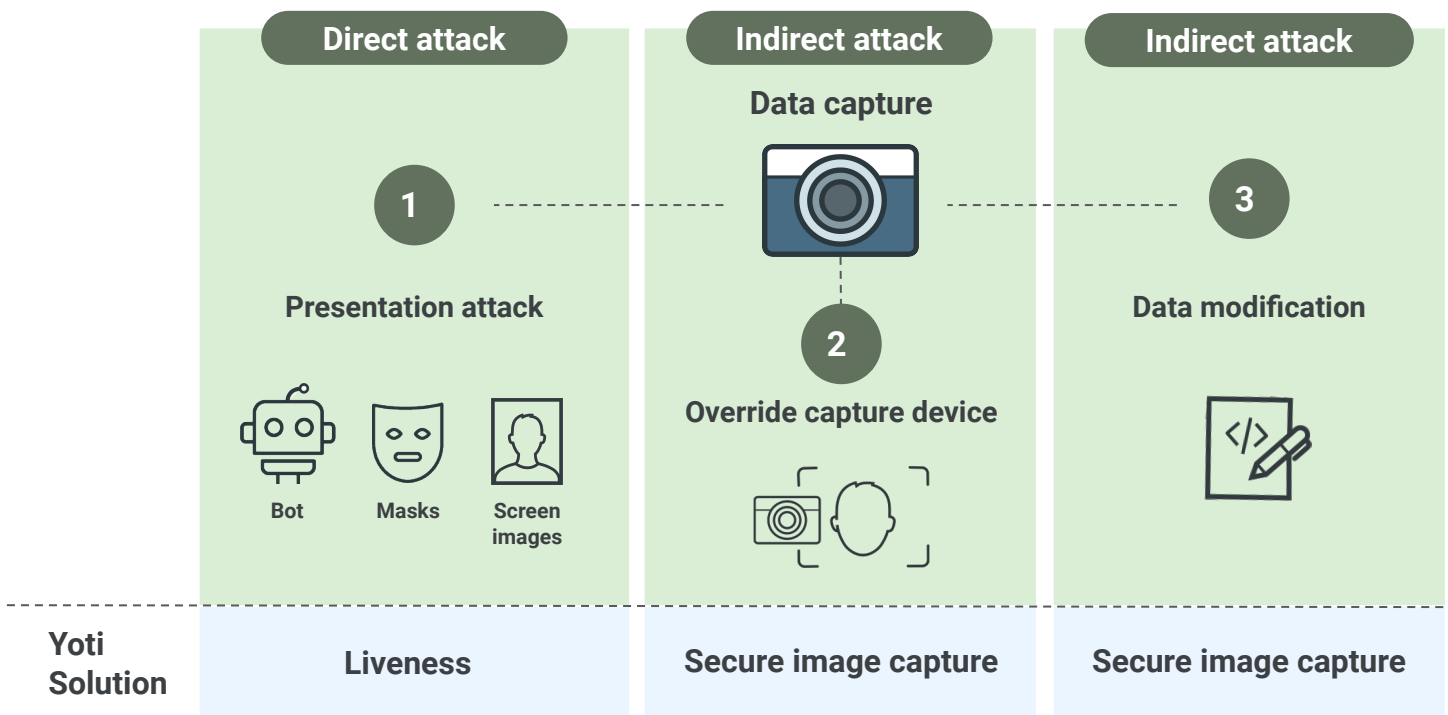
- Issue: Platforms may struggle to effectively moderate and detect AI-generated content.
- Threat: This could result in the spread of harmful or malicious content, as platforms may face difficulties in identifying and removing generated content.

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. That is, preventing an image or video from being used or uploaded at the point of verification.

Bad actors can 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

- **Are a relatively mature and well understood issue across the verification space**
- **Present an alternative image, video or mask to a live camera**
- **Are a more naive and more obvious attempt to spoof a verification check**

A presentation attack is an attempt to spoof a verification check 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 being verified is a real person. An attempt to spoof a verification in this way is often termed a 'direct' attack.

Yoti's proprietary passive liveness detection technology, MyFace® Live, is compliant with iBeta ISO PAD Level 2. The technology achieved a 100% attack detection rate during testing by NIST. MyFace® Live is a passive solution and only requires a selfie image to complete the liveness check. This creates a smooth, inclusive and accessible user experience.

We also offer businesses the option to use 2 other leading liveness suppliers that offer a fall back and complement to Yoti's proprietary system. This can be included as part of the process as either a 'double check', a fall back or for where 'high risk' has been identified.

Read more about liveness and MyFace® Live [here](#).

Injection attacks

- **Are an increasing threat vector**
- **Bypass the device camera to 'inject' an alternative image or video**
- **Only require tech understanding and investment to attempt an attack**
- **Are an attempt to overcome liveness checks**

An injection attack is an 'indirect' attack which 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** - these 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 has developed MyFace® SICAP (**Secure Image CAPture**), a patented 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. MyFace® SICAP can detect both software and hardware attacks.

There are two parts to how MyFace® 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.

By default we also block virtual camera software such as ManyCam, VCam, FakeWebCam.

By using both MyFace® Live and MyFace® SICAP products, Yoti solutions provide the latest technology available to combat attempts to spoof or produce generative AI content.



Developments - MyFace® SICAP 2.0

Our latest update, MyFace® 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 allow organisations to choose whether or not they receive the image. This is better for user privacy and data minimisation. For example, if an organisation requires an age check, there is no need for them 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 MyFace® 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 well understood generative AI models have been developed by researchers or benevolent commercial organisations with the goal of tricking the human eye - this 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, the situation 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 creating generative AI videos.



How it works

Let's look at three examples:

- 1) Intimate images
- 2) High-profile celebrities and public figures
- 3) Verification and authentication for account holders

Intimate images

There are multiple companies that are able to detect if an image or video contains explicit imagery. This can be automatically flagged and trigger a requirement to gain consent from the account owner to ensure they are the person in the image. This is the part Yoti can help with. We provide technology to match the person uploading the content with the image - ensuring it is the same person in the intimate image. We can then capture consent for all participants in a particular image or video.

In genuine cases, consent can be easily provided using our trust and safety tools. Where it's a deepfake image, the individual would not provide their consent. This would prevent the spread of misinformation and non-consensual, intimate imagery of real people.

Any suspicious content can be stopped at source. This would automate the process, tackling deepfakes at scale and reducing pressure on human moderation teams.

High profile celebrities and public figures

For celebrities and well known public figures, their profile can be added to a 'watch list', either by the platform to conduct their own checks, or at the individual's request. Any content uploaded which features their image can be immediately flagged as potentially problematic, requiring a further check.

This can work in many ways:

- Yes, this person has consented - no further checks are needed
- Yes, the content is from their verified profile - no further checks are needed
- We don't recognise this person - need to get consent before the content is uploaded
- No, this person is banned from the platform - the content needs to be removed
- No, this person is potentially underage - the content needs to be removed



By verifying content upon upload, deepfake images and videos can be prevented from appearing online. These checks can also be completed during live streaming - ensuring there is a real person in the content.

This could easily extend to a wider audience too. Members of the wider population could 'opt-in' to a service that would ensure no graphic or AI-generated content could be submitted to online platforms that participated.

This could all be achieved using a 1:N or 1:many facial recognition product such as Yoti MyFace® Match.

Verification and authentication for account holders

Another issue is deepfakes being used for high-value transaction authentication, or account takeovers. Firstly, the identity verification process to set up an account has been fairly well understood, transitioning from in-person, in-branch document checks and document signing to well established online processes involving AI technology and 'remote' human checking.





The emerging risk is around account takeovers and high-value transactions, where the potential reward for bad actors can be very high. Therefore the resources applied can be significant. Here, despite layers of security, deepfakes or injection attacks can override current systems, particularly where an account holder's personal information has been compromised. There is also the growing threat of SIM swaps, where a bad actor can take over a user's mobile phone account, thereby comprising an additional layer of authentication.

One solution can be to conduct a face match check at the point of transaction. A biometric map of the user's face can be stored either by the organisation or the IDSP and taken at the point of account setup. This can then be used to check a user is who they say they are when required. This would also use our anti-spoofing tools of liveness and SICAP to ensure the user is who they say they are and not a presentation or indirect attack, such as replay attacks or injection attacks.

Another quick and simple way of confirming is using a digital ID. For marketplace transactions or high value account withdrawals, the controller can simply request a 'peer 2 peer' share from the other party, confirming their name and email address, and within seconds can receive a verified confirmation.

Example use cases

The rise of generative AI, while offering numerous benefits and advancements in various fields, also brings forth certain potential threats and challenges. Some notable concerns associated with generative AI include:

- | | |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  <p>Dating</p> <p>Profile picture matching</p> |  <p>Financial services</p> <p>Strong account authentication and identity verification</p> |
|  <p>Gaming</p> <p>Ensure age appropriate experiences and enable parental consent</p> |  <p>Content moderation</p> <p>Verify content</p> |
|  <p>Social media</p> <p>Verify profiles and ensure content</p> |  <p>Marketplaces</p> <p>Reduce fraudulent buyers and sellers to increase trust and safety for your platform.</p> |

Alternative approaches

There are other organisations and methods of offering generative AI detection but we have not yet been able to verify their accuracy. It was also recently 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](#)). Referring back to our strategy, the pace of generative AI innovation can quickly make any generative AI detection quickly redundant.

Hence our strategy and approach of targeting the problem at source, as we've described above.



To learn more about Yoti's approach to combating
Generative AI please [get in touch](#).