Deepfakes on YouTube are on the rise: How to flag certain AI content?

Artificial intelligence (AI) is starting to impact our lives in tangible ways, both positive and negative.

And this forces tech companies to take steps to protect users of their platforms. YouTube is no exception. The platform now allows individuals to report videos with AI-generated content.

How are AI-generated YouTube videos on the rise?

AI-generated content is everywhere. News reports and online articles can be created by AI bots like ChatGPT and Gemini, photos and artwork can be created by DALL-E and Imagen, and videos can be created by a variety of tools. Text-to-video AI tools are improving.

Of the different forms of content that AI is capable of generating, video is currently the least refined, with numerous examples of AI models creating extremely bizarre and disturbing clips. However, technology is improving and simpler forms, such as putting words in someone's mouth, are becoming more difficult to detect.

AI-generated videos range from innocuous parodies of celebrities saying things they haven't actually said to fake videos designed to spread misinformation. This is being used to advance agendas, destabilize democracies and sway elections.

Either way, as technology improves and the number of AI-generated videos increases, online video sites and social media platforms will have to put up barriers to prevent abuse. This is why, as first reported by TechCrunch, YouTube is taking action to address the problem.

How to flag AI-generated content in YouTube videos

YouTube has quietly rolled out an option under its privacy guidelines for individuals to report videos that use AI-generated or synthetic characters that look or sound like you. This applies whether the video is intended to cause harm (such as ruining someone's reputation) or is intended to be more harmless (but still undesirable).

If you spot an artificial or AI-generated version of yourself in a YouTube video, you can now submit a takedown request. This is done by following the Privacy Complaints Process, which relies on the affected individual reporting the issue, unlike the ability of anyone to flag YouTube videos for other forms of abuse .

- **Report AI-generated or other synthetic content that looks or sounds like you**: if someone has used AI to alter or create synthetic content that looks or sounds like you, you can ask for it to be removed. In order to qualify for removal, the content should depict a realistic altered or synthetic version of your likeness. We will consider a variety of factors when evaluating the complaint, such as:
  - Whether the content is altered or synthetic
  - Whether the content is disclosed to viewers as altered or synthetic
  - Whether the person can be uniquely identified
  - Whether the content is realistic
  - Whether the content contains parody, satire or other public interest value
  - Whether the content features a public figure or well-known individual engaging in sensitive behaviour such as criminal activity, violence or endorsing a product or political candidate

To start a report, follow the Privacy Complaint Process for altered or synthetic content.

Learn more about altered or synthetic content labels.

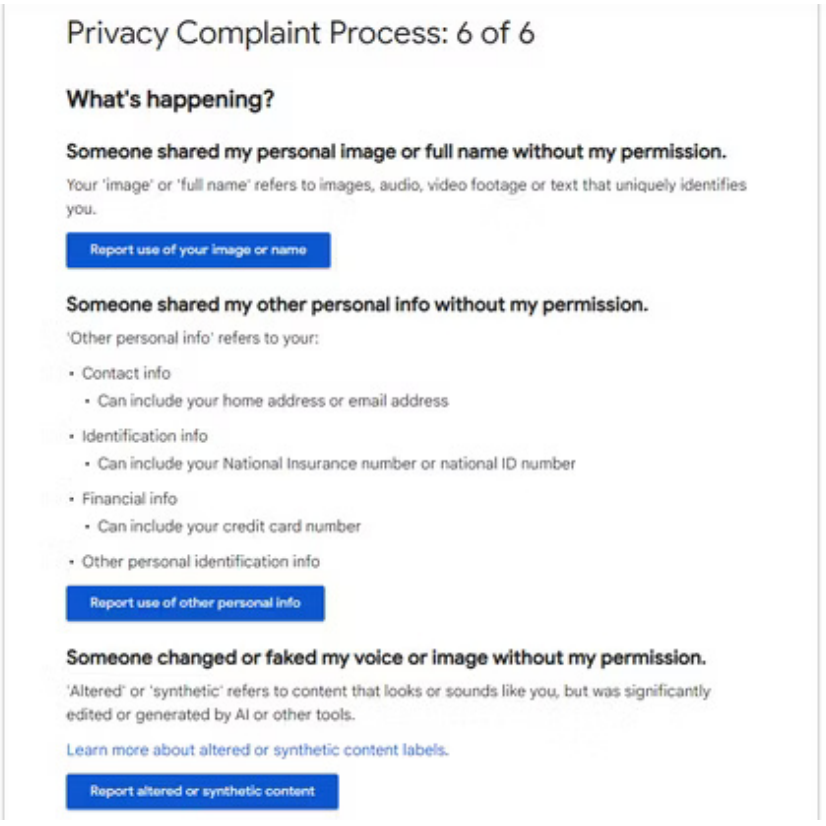
More about reporting a privacy violation

We take your privacy seriously. When you report a privacy violation, we will never share your name or contact information with the person who posted the content without your permission. We'll send updates to the email address that you provide in the complaint form.

First-party claims are required.

We do not accept claims made on behalf of third parties **except** in the following situations:

- The individual whose privacy is being violated does not have access to a computer or the Internet
- The individual whose privacy is being violated is a vulnerable individual
- The claim is being made by the parent or legal guardian of the individual whose privacy is being violated
- The claim is being made by a legal representative of the individual whose privacy is being violated
- The request is filed by a close family member on behalf of a deceased individual



To report a video that contains an AI-generated or other synthetically altered likeness of you, go through YouTube's Privacy Complaint Process. When you get to step 6 of 6, you'll be able to "Report altered or aggregated content". YouTube has an explanation page, briefly describing them as "content that looks like you but has been edited or generated using AI or other tools."

What happens when you flag AI content on YouTube?

After you follow YouTube's privacy complaints process and report altered or aggregated content, the YouTuber responsible will have 48 hours to remove the video. If the video is not removed within that time frame, YouTube will investigate further.

YouTube users also have the option of removing personal (and identifying) information from videos or blurring the faces of people involved in the video. However, they can't just make the video private to remove the claim because YouTube doesn't trust that the uploader won't change the status back to public later.

YouTube does not promise to remove all AI-generated content reported by the individual concerned, instead promising to *"consider a variety of factors when evaluating the complaint"* . The platform also requires first-party complaints (except in certain cases), meaning you won't be able to complain on behalf of a celebrity.

However, this is a remarkable first step in what is sure to become an increasing problem over the next few years.