Deepfake trends and challenges
- position statement

A deepfake is a digital photo, video or sound file of a real person that has been edited to create an extremely realistic but false depiction of them doing or saying something that they did not actually do or say. Deepfakes are created using artificial intelligence software that currently draws on a large number of photos or recordings of the person to model and create content.

Background

Manipulation of images is not new, but over recent decades digital recording and editing techniques have made it far easier to produce fake visual and audio content, not just of humans but also of animals, machines and even inanimate objects.

Advances in artificial intelligence (AI) and machine learning have taken the technology even further, allowing it to rapidly generate content that is extremely realistic, almost impossible to detect with the naked eye and difficult to debunk. This is why the resulting photos, videos and sound files are called 'deepfakes'.

To generate convincing content, most current deepfake technologies require large amounts of genuine data (images, footage or sound recordings). But technical capability is advancing quickly, so soon only a small number of original files will be needed to create a deepfake. Indeed, the field is evolving so rapidly that deepfake content can be generated without the need for any human supervision at all, using what is called recycled generative adversarial networks (commonly referred to as GANs).

Deepfakes have numerous positive applications in entertainment, education, medicine and other fields, particularly for modelling and predicting behaviour. However, the possibilities for abuse are growing exponentially as digital distribution platforms become more publicly accessible and the tools to create deepfakes become relatively cheap, user-friendly and mainstream.

Deepfakes have the potential to cause significant harms. To date, they have been used to create fake news, false pornographic videos and malicious hoaxes, usually targeting well-known people such as politicians and celebrities. Potentially, deepfakes can be used as a tool for identity theft, extortion, sexual exploitation, reputational damage, ridicule, intimidation and harassment.

A person who is targeted may experience financial loss, damage to professional or social standing, fear, humiliation, shame, loss of self-esteem and reduced confidence. Reports of misrepresentation and deception could undermine trust in digital platforms and services, and increase general levels of fear and suspicion within society.

Recent coverage

As advances in deepfake technology gather pace, and apps and tools are emerging that allow the general public to produce credible deepfakes, concerns are growing about the potential for harm to both individuals and society.

As noted in a blogpost by eSafety Commissioner Julie Inman Grant, Fighting the deepfakes arms race (October 2019), the development of innovations to help identify deepfakes is not yet keeping pace with the technology itself. More companies could be testing blockchain technology, digital watermarking, digital signatures and other tracking and tracing tools as deepfake detection solutions. However, using deepfakes to target and abuse others is not simply a technology problem. It is the result of social, cultural and behavioural issues being played out online. The Australian Strategic Policy Institute's recent report, Weaponised deepfakes (April 2020), highlights the challenges to security and democracy that deepfakes present — including heightened potential for fraud, propaganda and disinformation, military deception and the erosion of trust in institutions and fair election processes.

The risks of deploying a technology without first assessing and addressing the potential for individual and societal impacts are unacceptably high. Deepfakes provide yet another example of the importance of Safety by Design to assist in anticipating and engineering out misuse at the get-go.

eSafety approach

A holistic approach is needed to counter the negative impacts of deepfakes. eSafety leads this approach in Australia, working with industry and users to address the issue.

Our work includes the following:

- Raising awareness about deepfakes so Australians are provided with a reasoned and evidence-based overview of the issue and are well-informed about options available to them.
- Supporting people who have been targeted through a complaint reporting system. Any Australian whose photo or video has been digitally altered and shared online can contact eSafety for help to have it removed.
- Preventing harm through developing educational content about deepfakes, so Australians can critically assess online content and more confidently navigate the online world.
- Supporting industry through our Safety by Design initiative which helps companies and organisations to embed safety into their products and services.
- Supporting industry efforts to reduce or limit the redistribution of harmful deepfakes by encouraging them to develop:
  - policies, terms of service and community standards on deepfakes
  - screening and removal policies to manage abusive and illegal deepfakes
  - methods to identify and flag deepfakes to their community.

Advice for dealing with deepfakes

Deepfake technology is advancing rapidly, but there are still some signs that can help identify fake photos and videos.

Check for:

- blurring, cropped effects or pixilation (small ox-like shapes), particularly around the mouth, eyes and neck
- skin inconsistency or discoloration
- inconsistency across a video, such as glitches, sections of lower quality and changes in the lighting or background
- badly synced sound
- irregular blinking or movement that seems unnatural or irregular
- gaps in the storyline or speech.

If in doubt, remember to question the context. Ask yourself if it is what you would expect that person to say or do, in that place, at that time.

Some platforms are identifying and labelling deepfake or 'manipulated' content to alert their users.

How eSafety can help

Australians whose images or videos have been altered and posted online can contact eSafety for help for help to have them removed.

eSafety investigates image-based abuse which means sharing, or threatening to share, an intimate photo or video of a person online without their consent. This includes intimate images that have been digitally altered like deepfakes.

We can also help to remove:

- online communication to or about a child that is seriously threatening, seriously intimidating, seriously harassing or seriously humiliating - known as cyberbullying
- illegal and restricted material that shows or encourages the sexual abuse of children, terrorism or other acts of extreme violence.

Find out more about reporting harmful online content to eSafety.

Published: 23 January 2022