IDEMIA helps the French government to build its own deepfake detection toolbox

French Government Collaborates with Idemia to Counter Deepfake Threats

# JUSTICE & PUBLIC SAFETY

POSTED ON 05.27.24

What is deepfake and how does it work?

Deepfake is a machine learning technology that manipulates or fabricates video and audio recordings of people doing and/or saying things that never actually happened. This category of deepfake media uses various technologies and was officially born in the fall of 2017, appearing on the Reddit website, which is an American social news aggregation, content rating, and forum social network. Since 2017, the number of deepfakes has increased significantly. According to researchers at Deeptrace, deepfakes almost doubled over a two-year period; in 2019, there were approximately 15,000 deepfake videos found online, compared to almost 8,000 videos identified in 2018.

Deepfakes appear authentic and realistic, but they are not. They violate the victim's privacy, rights, and public image. Deepfakes can be used for criminal purposes: identity theft, attempts to discredit a person's reputation, disinformation campaigns, security fraud, extortion, online crimes against children, crypto jacking, illicit markets, etc.

Deepfakes have the potential to cause grave individual and societal harm and provoke violence all over the globe with just a click. They can exploit and magnify distrust in politicians, business front-runners, and other influential leaders. Furthermore, the pursuit of truth is on the line as well. Technology providers expect that advances in AI may make it difficult, if not impossible, to tell the difference between a deepfake and authentic media.

For these reasons, a solution is needed from tech companies, lawmakers, law enforcers, and the media—they all play a key role in helping keep society safe and protecting an individual's privacy. Politicians currently face the biggest danger, as they are often being filmed for various reasons, making them an easy and impactful target for deepfakes.

Despite high expectations and numerous research projects, debunking deepfakes is still far more difficult than generating them.

A Prototype Assessment Toolbox for Forensic Experts

In 2019, the French Service National de Police Scientifique (SNPS) identified the need for forensic tools that are resistant to modern deepfake generation methods.

In 2022, the French National Research Agency (ANR) launched a project called A Prototype Assessment Toolbox for Forensic Experts (APATE) to address this need. Five organizations are involved in this project which addresses the need for innovative deepfake detection methods as well as reliable and explainable tools for forensic experts, so that deepfake cases can be brought to court.

The research by the ANR includes:

- keeping deepfake generation and detection methods up to date.
- gathering or producing deepfakes for learning and testing.
- deepfake detection on images and image sequences using:
    - the effect of deepfake on low-level traces of noise, blur, compression, etc.
    - temporal inconsistencies with fully automated or self-supervised methods.
- deepfake detection on audio using:
    - speaker recognition techniques.
    - anti-spoofing techniques.
- deepfake detection, relying on the consistency of audio and image features.
- development of a deepfake detection toolbox, usable in court by forensic experts.

Today, experts rely on image falsification detection tools. The goal of APATE is to provide forensic experts with criteria and associated tools (scores, statistical distribution, masks applied on images to show areas of manipulation) and knowledge, enabling them to make objective decisions. Ideally, these tools should be adaptable and able to evolve to detect future deepfakes.

Benefits

APATE will help combat negative impacts at a:

- societal level (damage to economic stability / the justice system, manipulation of elections, etc.).
- psychological level (intimidation, defamation, etc.).
- financial level (extortion, brand damage, etc.).

It will open the door to legal procedures for deepfake-related crimes.

From an operational point of view, APATE will support law enforcement agencies by:

- improving agencies' investigative capabilities by providing forensic experts with a new set of tools with more accurate results. Past cases show that it is often difficult for experts to give a positive or negative answer when it comes to deepfakes due to a lack of reliable and accurate tools.
- providing an up-to-date and evolving toolbox.

APATE will also have a great scientific impact on the research community, notably by fostering research in multimodal and cross-modal representation and building expertise in speaker and face recognition to tackle counterfeiting issues.

IDEMIA's role in APATE

The SNPS needs a tool that can detect deepfakes and whose results can be presented as evidence in a court of law. To help the SNPS with this challenge, IDEMIA assembled a consortium consisting of SNPS, LRE-EPITA, École Polytechnique (l'X), ENS Paris Saclay, and IDEMIA, itself, and created a project for deepfake detection. Each partner is a leader in their field, making the consortium a complementary partnership.

This project was then presented to the ANR, who selected it because it provides a pertinent solution to one of the problems the ANR had identified. In addition, identification of the SNPS, as an end user, reinforced the ANR's choice.

From the outset, IDEMIA has ensured that the SNPS is included in all decisions. SNPS' role is to test, challenge, and implement the tool. It will evaluate the tool and solidify elements of the research by exploring image and video deepfakes and providing an algorithm to detect them.

APATE's status and the way forward

The project started in September 2022 and is currently in its preliminary stages. Until now, the project has been able to catalog deepfake attacks. It has also provided an exhaustive view of all techniques used to date.

The goal, now, will be to learn about the different deepfake detection models. This will be done using the inventoried databases and the compiled list of state-of-the-art deepfakes, which will enable the creation of the first toolbox.

Mandatory mention

This research program has been partially funded by the French National Research Agency (ANR). under project code « ANR-22-CE39-0016-05 » and through an Advanced Grant from the European Research Council (ERC).

This article explores IDEMIA's collaboration with the French government in developing a deepfake detection toolbox, as part of the APATE project funded by the French National Research Agency (ANR). It discusses the rising threat of manipulated media and the efforts to combat it through innovative forensic analysis methods.