Pressemitteilung

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, DFKI
Jeremy Gob

11.06.2024
http://idw-online.de/de/news835036

Forschungs- / Wissenstransfer
Informationstechnik, Kulturwissenschaften, Medien- und Kommunikationswissenschaften, Psychologie, Sprache / Literatur
überregional

Targeting deepfakes: AI as a weapon against digital manipulation

Deepfakes represent a serious challenge that raises both technological and societal questions. Researchers at the German Research Centre for Artificial Intelligence (DFKI) in Berlin are therefore developing methods to reliably detect deepfakes in order to reach people with the necessary warnings and corrections. ‘News-Polygraph’ is the name of the ambitious project that aims to gain a decisive advantage in the cat-and-mouse game between the products of generative ai-models and recognition technologies.

Deepfakes are realistic-looking media content that is created or manipulated using generative artificial intelligence (genAI) to generate deceptively real audio, video and image content. The possible applications: Almost limitless! Vera Schmitt, guest researcher from TU Berlin at DFKI Berlin, and Tim Polzehl, DFKI researcher in the ‘Speech and Language Technology’ department, shed light on how this technology unfolds its positive and negative potential and how we as a society can protect ourselves from disinformation and manipulation by providing insights into their work.

Tim Polzehl, researcher in the Speech and Language Technology department at DFKI: "We know an early version of what is considered a deepfake today from speech synthesis. There, AI is used to develop computer-generated voices that sound as real as possible, which have developed to such an extent in the last five years that individual voices can now be generated deceptively realistically - even with little training material. Today, generative AI also enables the creation of deceptively real images, videos and audio that are often difficult to distinguish from real content. With the rise and public availability of generative AI, the topic has become a broad social phenomenon that raises technical, ethical and application-related questions."

These questions demand answers. As researchers, Tim Polzehl and Vera Schmitt are looking at how technology can help answer these questions. However, successfully identifying manipulative AI-generated media content requires not only technical but also human solutions. A circumstance that already poses a challenge in the definition of ‘deepfakes’.

Vera Schmitt, guest researcher in the Speech and Language Technology department at DFKI: "It is difficult to find a single exact definition for “deepfakes” - there are a multitude of definitions. Deepfakes are basically realistic media content that is altered, generated or falsified by AI and Transformer-based models. However, there is a big debate to be had about the extent to which intent, fraud, blackmail, damage to reputation and political manipulation play a role, and to what extent art and entertainment should be given a place."

Humans and AI: strong together

People and technology need to think together. After all, it is people who believe, process impressions, are manipulated - and possibly manipulate themselves. Only through the combination of human judgement and AI-supported tools can we reliably recognise when a deception is taking place and thus develop effective countermeasures.

This interaction is necessary because intent to deceive cannot be recognised well by AI models, for example, as this requires suitable indicators. Humans must therefore control the evaluations of AI models, create the context themselves and consider other possibilities.

We humans recognise certain indicators of counterfeiting straight away, while AI recognises other signs.

It's all about the details

A striking example: Let's look at a realistic-looking photo of a person who has two earlobes on each side. For most people, this would be a clear sign of a deepfake.

Tim Polzehl: "The AI stumbles at this point, because all the elements recognised by the AI may look realistic - and the earlobe may not be recognised at all. Or it is recognised, but the AI cannot put what it has recognised into a meaningful context - namely that we humans generally only have one earlobe per ear. In order to make such a decision, an AI would first of all need significantly improved earlobe recognition, logical, critical and questioning thinking and global knowledge of human anatomy, which is not currently available. We humans have these abilities and can deduce from our knowledge and the context that this photo is probably not authentic."

When it comes to image representations and technical subtleties, however, AI is way ahead of us. Lighting conditions, shadows and overlays, movements, transitions and anomalies at pixel level - these are the areas where it becomes difficult for human perception. AI tools can provide excellent assistance with this almost forensic observation, because highly specialised AI works very well - in other words, it can perform explicit tasks effectively. Irregularities and anomalies can then be interpreted as indicators of content generated using generative AI models.

Content-based analysis

Apart from identifying inconsistencies in content, humans are able to incorporate proportionalities and expectations into their consideration of media content.

Vera Schmitt: "In most cases we have a good understanding of context and logic. So if the pillars at the Brandenburg Gate topple over in a video and people standing around don't react to this event at all, then we can very easily conclude that it is a fake representation. In addition, there would be a large number of independent sources reporting on such an event."

In order to identify deepfakes and manipulative content, it is therefore necessary to analyse the content. Especially if the form of presentation makes it difficult to distinguish between authentic and artificial material - as is the case with text, for example.

There are now many different popular transformer-based models for text-generation. These synthetic textual products are almost impossible to recognise in small quantities. Both for humans and for AI.

Vera Schmitt: "This is why answering central questions is fundamental to recognising false information. Who originally circulated the information? What facts, people and events are being presented? Are there already known fakes on the subject?"

Specialised AI tools can already provide reliable answers to these questions. Publicly available applications such as Deep Ware Scanner, Deeptrace or Wisper can be used to validate information. In future, the news polygraph should also empower people to check information more easily - and uncover manipulative narratives.

Tim Polzehl: "We are dealing with two concepts. Firstly, there are deepfakes, i.e. audio, videos, images and the like with supposed authenticity. Then there is disinformation in narratives. The latter brings us into the area of fact-checking - and to our news polygraph."

News polygraph vs. disinformation in narratives

A basic idea of fact-checking is that manipulative narratives repeat themselves, so we can look into the past - and possibly discover the same narratives again in the present. AI can successfully support this process. We then need to check whether the narrative has already been refuted, whether it has already been published - and finally, how this information can be communicated effectively.

Polzehl and Schmitt's team views their news-polygraph as an ‘AI model for intelligent decision support for journalists’. It is therefore crucial that the analyses of the model can be presented in such a transparent way that journalists can understand and categorise them accordingly.

Vera Schmitt: "It is also important to evaluate the circumstances surrounding the spread of misinformation and disinformation and to incorporate these into an assessment or implementation, such as a Digital Service Act. After all, fake content can also be shared unknowingly and unintentionally, without any intention to deceive."

A procedure would therefore be needed to not only label AI-generated material, but also to measure its intention and impact in addition to its authenticity. The fact that AI can generate synthetic media such as voices, videos and images is initially positive, says Schmitt. However, people can run personal campaigns with the same content and misuse these media.

Tim Polzehl: "Arming yourself against disinformation therefore means questioning more often and more critically who and, above all, why you believe certain claims. The intention and sources of a claim play an increasingly important role. This also applies to us scientists. If, for example, communication is based on facts, sources are usually also provided. In the end, however, we scientists also have to give away our trust to some extent - even if science is subsequently largely based on evidence."

There is no such thing as absolute certainty

Vera Schmitt: "There will never be an AI that can recognise everything. Furthermore, there is an immense imbalance between generative models and recognition technologies, which needs to be balanced out by an increase in resources and attention for this topic. Because deepfakes have an almost infinite reach in today's world - a scalability - which must be countered by education, relief and empowerment."

This is another reason why Polzehl and Schmitt shared their assessments at this year's re:publica in Berlin. However, even if information and a critical approach in combination with AI tools enable deepfakes to be recognised more reliably in the future, this will not defuse them. The dangers lie behind the artificially created façade.

Tim Polzehl: "Even the labelling of AI-generated material does not necessarily protect against being influenced by this content! A study on labelling revealed that people can still be influenced. So ‘recognising’ doesn't mean the issue is off the table. It is my personal wish that we, as a society, better categorise the importance of disinformation so that we are better prepared for it. Then labelling can work, harmful narratives and content can be successfully intercepted and better monitoring can be carried out. All of this should happen simultaneously to relieve the burden on all people who are consistently confronted with a growing number of counterfeits. And enable dedicated actors to cope with the growing output."

wissenschaftliche Ansprechpartner:

Vera Schmitt, guest researcher in the Speech and Language Technology department at DFKI
Vera.Schmitt@dfki.de

Tim Polzehl, researcher in the Speech and Language Technology department at DFKI
+49 30 23895 1863
Tim.Polzehl@dfki.de

Originalpublikation:

https://www.dfki.de/en/web/research/projects-and-publications/project/news-polygraph

URL zur Pressemitteilung: https://www.youtube.com/watch?v=vduo1kYlHgk&t;=1s

Deepfakes - our new reality
DFKI/Midjourney