Fälschung und Manipulation

Durch künstliche Intelligenz (KI) wird es zunehmend einfacher, Foto-, Video- oder Audiodateien zu manipulieren. Seit 1. Jänner 2024 werden Deepfakes von der Polizei systematisch erfasst. Ermittlungen können dadurch zielgerichtet geführt und die Präventionsarbeit verbessert werden.

Die Möglichkeiten, mediale Identitäten zu manipulieren, existieren bereits seit vielen Jahren. Dass Bilder durch vielfältige Methoden ver- oder gefälscht werden können, ist spätestens seit Spaß-Apps, die sogenanntes „Face-Swapping“ auf simple und schnelle Art und Weise ermöglichen oder diversen Challenges auf sozialen Netzwerken, wie etwa der #FaceAppChallenge, der breiten Öffentlichkeit bekannt. Was auf den ersten Blick wie ein harmloser Spaß aussieht, entwickelte sich in den vergangenen Jahren durch den Einsatz von KI zu einer erheblichen Gefahr. Mittlerweile können Deepfakes zunehmend einfacher erstellt und auch missbräuchlich genutzt und weniger leicht erkannt werden. Sie können zur politischen Einflussnahme, Verbreitung von Fake News oder Begehung von Straftaten eingesetzt werden.

Deepfake-Betrug: Die vermeintlichen Teilnehmer einer Videokonferenz sollen größtenteils KI-generierte Nachbildungen von echten Menschen gewesen sein

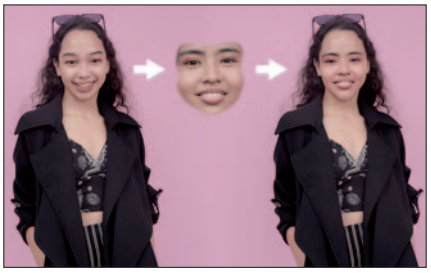
Gefakte Videokonferenz. Für Aufsehen sorgte ein Betrugsfall Anfang Februar 2024, bei dem ein multinationales Unternehmen in Hongkong um umgerechnet rund 23 Millionen Euro geprellt worden sei. Dabei sei ein Angestellter des Unternehmens nach einer zunächst per E-Mail erfolgten Zahlungsaufforderung vom vermeintlichen Finanzchef zu einer Videokonferenz eingeladen worden sein. Die vermeintlichen Teilnehmer der Videokonferenz sollen größtenteils KI-generierte Nachbildungen von echten Menschen gewesen sein.

Um gegen diese Entwicklung vorzugehen, wurde der nationale Aktionsplan gegen Deepfakes ausgearbeitet und im Mai 2023 präsentiert. Eine wichtige Maßnahme wurde mit 1. Jänner 2024 nun umgesetzt: Deepfakes werden von der Polizei systematisch erfasst.

Deepfakes sind Fotos, Videos oder Audiodateien, die mit Hilfe von KI verändert werden. Der Terminus setzt sich aus den Begriffen „deep learning“ (eine Methode, durch die eine KI lernt) und „fake“, also Fälschung bzw. Verfälschung zusammen. Deepfakes werden als Überbegriff für verschiedene Formen von audiovisueller (medialer) Manipulation verwendet, bei denen deren Echtheit bzw. Manipulation mit dem bloßen Auge nicht mehr bzw. nur schwer feststellbar ist.

Fälschung von Gesichtern. Um Gesichter in Videos zu manipulieren, wurden in den vergangenen Jahren mehrere KI-basierte Verfahren entwickelt. Ziel ist es, entweder die Gesichter im Video zu tauschen (Face-Swapping), die Mimik bzw. Kopfbewegung einer Person im Video nach Wunsch zu verändern (Face-Reenactment) oder neue Identitäten zu erstellen, die nicht existieren.

Durch das Face-Reenactment ist es möglich, täuschend echte Videos zu produzieren, in denen Personen Aussagen in den Mund gelegt werden, die in der Realität nie getätigt wurden. Dafür wird anhand eines Videostreams ein 3D-Modell des Gesichts der Zielperson erzeugt. Dieses Modell kann der Erzeuger dann beliebig mit seinem eigenen Stream kontrollieren und eine echt aussehende Mimik generieren.



„Face-Swapping“: Ersetzen eines Gesichts in einem Bild durch das einer anderen Person

Fälschung von Stimmen. Um manipulierte Stimmen zu erstellen, werden insbesondere „Text-to-Speech (TTS)“- und „Voice-Conversion (VC)“-Verfahren eingesetzt. Beide Methoden haben das Ziel, einer Person eine Aussage zuzuschreiben, die sie nie getätigt hat. Mit der TTS-Methode ist es zudem nicht nur möglich Menschen in die Irre zu führen, sondern auch automatisierte Sprecherkennungsverfahren zu täuschen. Damit diese Verfahren funktionieren, muss die KI zunächst mit Trainingsdaten „angelernt werden“, wobei sich die Art der notwendigen Daten je nach Angriffsart und -ziel unterscheidet. Alle Methoden haben jedoch gemeinsam, dass von der Zielperson Audio-Aufnahmen in möglichst konstant hoher Qualität benötigt werden.

Verschiedenste Bedrohungsszenarien. Dadurch, dass die Methoden bereits von technisch versierten Laien einsetzbar sind, können Deepfakes missbräuchlich verwendet werden. Ob bei gezielten Phishing-Angriffen, Desinformationskampagnen oder Verleumdungen: Deepfakes können als KI-basierter

Fotos: Adobe Stock.com

Modus Operandi großen Schaden anrichten. Laut Sensity-Studie „The State of Deepfakes“ aus dem Jahr 2020 kommen Deepfakes vor allem im pornografischen Bereich zum Einsatz. Bei 93 Prozent aller Deepfake-Videos, die online zu finden waren, waren die Inhalte pornografischer Natur, die ausschließlich auf Frauen abzielen und ihnen schaden sollen. Um einem Missbrauch der eigenen Bilder entgegenzuwirken, muss hinterfragt werden, welche Inhalte mit der Öffentlichkeit geteilt werden. Schwieriger ist es hierbei für Menschen des öffentlichen Lebens, wie eine berühmte US-Sängerin Ende Jänner 2024 erfahren musste. In sozialen Netzwerken wurden angebliche Nacktbilder der Künstlerin veröffentlicht und nur durch das schnelle Einschreiten der Betreiber konnte der Verbreitung der Deepfake-Bilder Einhalt geboten werden.

Protokollierung seit 1. Jänner 2024. Deepfake-Videos oder -Audios können aus kriminalpolizeilicher Sicht zur Verwirklichung zahlreicher strafrechtlicher Delikte verwendet werden, wie Betrug, Erpressung, gefährliche Drohung, Online-Kindesmissbrauch etc. Um Straftaten mit diesem Modus Operandi strukturiert erfassen und auswerten zu können wurde mit 1. Jänner 2024 der Code „Deepfake“ im Protokollierungssystem der Polizei (PAD) als neue Begehungsform eingeführt. Wenn jemand in einer Polizeiinspektion eine Anzeige erstattet wegen des Verdachts, Opfer einer Deepfake-Manipulation geworden zu sein oder eine solche meldet, wird das im PAD protokolliert. Dadurch kann zukünftig ein umfassendes Lagebild über die Verbreitung des Modus Operandi erstellt werden und die Ermittlungen zielgerichtet geführt werden.

Digitale Spuren. Jeder hinterlässt digitale Spuren und so können auch bei der Manipulation von Medieninhalten etwa Metadaten in den Dateien, Artefakte von Bild- und Videobearbeitungswerkzeugen oder Unregelmäßigkeiten in den Pixeln von den Multimedia-Forensikern des Cybercrime-Competence-Centers (C4) des Bundeskriminalamtes ausgelesen und analysiert werden. Sollte ein Deepfake-Verdacht bestehen, können die einschreitenden Beamtinnen und Beamten die zu untersuchenden Dateien dem C4 übermitteln.

Wie erkennt man Deepfakes? Sichtbare Übergänge, etwa an der Naht rund um das Gesicht, verwaschene Konturen, wie bei Zähnen oder Augen, unnatürliche Mimik, unlogische Schatten oder Hintergründe und ein fehlendes Blinzeln können Hinweise für eine Gesichtsmanipulation sein. Typische Artefakte bei synthetischen Stimmen können etwa ein metallischer Klang, eine falsche Aussprache oder Sprechweise oder auch eine monotone Sprachausgabe sein.

Tools. Unterschiedliche Unternehmen und Universitäten arbeiten zudem an Tools, um Deepfakes zu erkennen. So können etwa der Scanner von *Deepware (Deepware | Deepware.Ai | Scan & Detect Deepfake Videos)* oder von *Deep-Fake-O-Meter (DeepFake-o-meter (buffalo.edu))* Bild-, Video- und Audiodateien überprüfen und einen Anhaltspunkt für den Wahrheitsgehalt der Datei liefern. Für BMI-Bedienstete steht am e-Campus der SIAK ein E-Learning-Kurs „Deepfakes“ bereit. Er vermittelt Basiswissen zum Thema und bietet Unterstützung für den Umgang mit manipulierten Medieninhalten an.

Romana Tofan