Christian Gerber

Deepfakes: **Wachsende Bedrohung** für die Finanzindustrie?

Sogenannte „Deepfakes" werden hauptsächlich zur Unterhaltung in der Filmindustrie, für politische Satire oder im pornografischen Bereich verwendet. Inzwischen häufen sich – nicht zuletzt auch durch die Corona-Pandemie – jedoch Fälle, bei denen Kriminelle die Technologie einsetzen, um Personen und Unternehmen zu manipulieren und zu betrügen. Führungskräfte und Compliance-Abteilungen sollten sich daher proaktiv mit dem Thema auseinandersetzen und durch ein besseres Verständnis der Technologie Lösungen implementieren, um sich und ihr Unternehmen davor zu schützen.

„Mit Deepfakes können Fälschungen mittels künstlicher Intelligenz weitgehend autonom erzeugt werden."

Der Begriff „Deepfake" setzt sich zusammen aus „Deep Learning", einer Methode für maschinelles Lernen, und dem englischen Wort „Fake" („Fälschung"). Er beschreibt realistisch wirkende Bild-, Ton- und Videoinhalte, welche abgeändert und/oder verfälscht werden, um zum Beispiel Gesichter in Videos auszutauschen (sogenannte „Face Swaps").

Die Manipulation von Medien ist an sich kein neues Phänomen. Mit Deepfakes können solche Fälschungen mittels künstlicher Intelligenz jedoch weitgehend autonom erzeugt werden. Die verwendete Software kann in der Regel eine leicht nutzbare Standardsoftware für Bildbearbeitung bis hin zu kostenlosen Open-Source-Anwendungen sein.

Eines der bekanntesten Beispiele für den betrügerischen Einsatz von Deepfakes ereignete sich im März 2019 in einem britischen Energieunternehmen. Dessen Chief Executive Officer erhielt in einem Telefonat mit dem Leiter der deutschen Muttergesellschaft den Auftrag, die Summe von 220 000 Euro binnen einer Stunde an einen Lieferanten im Ausland zu überweisen. Tatsächlich gehörte die Stimme jedoch einem Betrüger. Ein Mitarbeiter, der zur Schadensregulierung beauftragten Versicherung Euler Hermes, berichtete gegenüber dem Wall Street Journal, dass der Betrüger künstliche Intelligenz (KI) nutzte, um ein Deepfake der Stimme des deutschen Managers zu erstellen.

Die Stimme sei dabei so perfekt nachgemacht, dass sie sowohl den subtilen deutschen Akzent sowie die Sprachmelodie des deutschen Vorgesetzten täuschend echt simulierte. Insgesamt rief der Betrüger wohl auf diese Weise insgesamt dreimal im Unternehmen an, um Überweisungen zu beauftragen. Erst beim dritten Anruf wurde der britische CEO skeptisch. Da war es jedoch zu spät. Die erste Überweisung war schon auf dem Konto des Betrügers angekommen

Erstellung und Verbreitung

Das niederländische Unternehmen Sensity, welches sich der Erkennung von Deepfakes verschrieben hat, durchsucht jährlich das Internet nach Deepfake-Videos. In dessen Deepfake Report 2019 berichtet es, dass fast 14 700 Deepfake-Videos im Internet gefunden worden seien. Dies entspreche einem Anstieg von über 100 Prozent im Vergleich zum Vorjahr. Allerdings seien 96 Prozent der gefundenen Videos lediglich pornografischen Inhalts gewesen. Diese Zahl sollte Unternehmen jedoch nicht in Sicherheit wiegen, da die wenigsten Kriminellen ihre Videos öffentlich ins Netz stellen. Eines ist sicher: Die Anzahl und Qualität der Deepfakes steigt stetig an. Weshalb der Einsatz von Deepfakes für kriminelle Aktivitäten dennoch (noch) nicht so einfach ist, liegt an der großen Zahl an Gesichtsaufnahmen, Sprachdateien oder Videos beider Personen, welche für einen täuschend echten „Face Swap" benötigt werden. Während man diese im Fall von Promi-

„Die wenigsten Kriminellen stellen ihre Videos öffentlich ins Netz."

nenten und öffentlichen Personen oftmals frei im Internet findet, ist dies bei Privatpersonen deutlich schwerer.

Hat man es dennoch geschafft, ausreichend Rohmaterial zu finden, kommt ein

sogenannter „Encoder"-Algorithmus zum Einsatz. Dieser findet und lernt Ähnlichkeiten zwischen den beiden Gesichtern und reduziert sie auf ihre gemeinsamen Merkmale. Darauf folgt die Bearbeitung durch das Gegenstück, die Decoder. Diese werden in vielen Wiederholungen darauf trainiert, das Gesicht einer der beiden Personen, basierend auf dem Modell des Encoders, wiederherzustellen. Als letzten Schritt kann man schließlich das Originalvideo von Person A an den Decoder übergeben. Dieser wendet dann sein erlerntes Wissen an und rekonstruiert Gesichtsmerkmale der falschen Person B im Gesicht der originalen Person A.

Selbst wenn man erste Ergebnisse mit diesen Methoden sehr schnell automatisch erzeugen kann, sind realistische Deepfakes, die selbst geschulte Augen täuschen, deutlich aufwändiger. Grund ist, dass das Rohmaterial nur in den seltensten Fällen in perfekter Qualität vorliegt und die Algorithmen daher insbesondere bei Konturen und feinen Details Fehler machen. Diese müssen für den „perfekten" Deepfake anschließend in mühsamer Kleinarbeit korrigiert werden.

„Es stellt sich die Frage, wie man manipulierte und künstliche Bilder und Stimmen erkennen kann."

Die Erschaffer eines bekannten, gefälschten Obama-Videos, in dem dieser sich selbst als Idiot bezeichnet, gaben an, dass ihr erfahrenes Produktionsteam 56 Stunden brauchte, um das einminütige Video zu produzieren.

Als weitere Möglichkeit neben dem Austausch von Gesichtern, kann man mit sogenannten „Generative Adversarial Networks" und einer großen Anzahl von Portraits die Deepfake-Algorithmen so trainieren, dass diese am Ende künstliche Gesichter erzeugen. Ein Beispiel hierfür ist die Webseite thispersondoesnotexist.com. Jedes Mal, wenn man diese Seite aufruft, wird ein täuschend echtes Foto einer Person generiert, die im echten Leben nicht existiert. Solche künstlichen Personen werden zum Beispiel im Bereich von Marketing und Forschung genutzt. Betrüger können sie jedoch auch verwenden, um falsche Identitäten zu erzeugen, die innerhalb herkömmlicher Onboarding-Prozesse nicht leicht zu erkennen sind.

Mit bloßem Auge kaum erkennbar

Es stellt sich natürlich die Frage, wie man manipulierte und künstliche Bilder und Stimmen erkennen kann. Ältere Deepfake-Bilder und -Videos und solche, die aus relativ wenig Rohdaten der jeweiligen Personen kreiert wurden, kann man oftmals durch eine manuelle Prüfung erkennen. Typische Merkmale sind unscharfe Stellen, sowie sichtbare Störungen und Artefakte im Bild. Typische Hinweise für Deepfakes sind unter anderem:

- Augen: Blinzelt jemand zu viel oder zu wenig?
- Haare: Wirken diese realistisch? Passen die Augenbrauen zum Gesicht?
- Haut: Ist der Hautton unregelmäßig und/oder wirken die Konturen falsch?
- Beleuchtung: Sind Reflexionen (zum Beispiel auf der Iris) widersprüchlich?
- Synchronisation: Bewegen sich die Lippen passend zum Gesagten?

Hierauf kann man sich jedoch nicht verlassen. Wie so oft, gibt es auch bei Deepfakes ein Katz-und-Maus-Spiel. Nimmt man zum Beispiel den ersten Punkt mit dem falschen Blinzeln. Dieser wurde im Jahr 2018 entdeckt. Den Grund hierfür erkannte man schnell. Die meisten Bilder zeigen Menschen mit offenen Augen und nur selten mit geschlossenen Augen. Somit hatten die Algorithmen zunächst kaum etwas über natürliches Blinzeln gelernt. Kaum war dies bekannt, wurden die Algorithmen speziell auf diese Merkmale trainiert, sodass heutzutage gute Deepfakes hier keine Fehler mehr aufweisen.

Drei Elemente für zusätzlich wirksamen Schutz

Für einen wirksamen Schutz gegen Deepfakes verwendet man heutzutage Computerprogramme, welche ironischerweise selbst mit künstlicher Intelligenz trainiert wurden, um Deepfakes zu erkennen. Dabei werden verschiedene Ansätze kombiniert. Einfachere Algorithmen können beispielsweise Artefakte auf der Ebene einzelner Bildpunkte (Pixel) erkennen oder anhand der Farben kleinster Bildausschnitte die Konsistenz mit dem Gesamtbild prüfen. Andere Ansätze bestehen in der Erkennung von Abweichungen in den Metadaten des Bildes (Angaben zur verwendeten Kamera, Brennweite und Belichtungsdauer) und dem tatsächlichen Bild. Aufwändige Verfahren nutzen räumliche oder zeitliche Analysemethoden. Sie werten zum Beispiel die 3D-Geometrie des Bildes aus und analysieren, ob es Unstimmigkeiten in der Kopfhaltung der Person gibt. Bei Videos vergleichen sie sämtliche Einzelbilder

und analysieren, ob die Bewegungen natürlich sind.

Je mehr man sich mit Deepfakes auseinandersetzt, umso mehr stellt sich für Unternehmen mit digitalen Geschäftsmodellen die Frage, wie man auf diese neue Bedrohung reagieren soll. Die Lösung ist jedoch denkbar einfach. Sie beruht auf der Integration digitaler Video- und Auto-Ident-Verfahren mit Anti-Spoofing-Funktionen in den Onboarding- beziehungsweise KYC-Prozessen. Diese sollten die folgenden Elemente enthalten, um einen wirksamen Schutz vor Deepfakes zu gewährleisten:

Prüfung der Sicherheitsmerkmale des ID-Dokuments: Aktuell ist im Rahmen des Video-Ident-Verfahrens von der BaFin lediglich vorgeschrieben, dass Anbieter zur Verifizierung des Dokuments drei optische Sicherheitsmerkmale prüfen. Während einige dieser Merkmale heutzutage recht gut gefälscht werden können, kann das Hologramm und dessen Reflektionen beim Kippen ein Schlüsselfaktor bei der Überprüfung der Integrität und Authentizität des ID-Dokuments sein. Für maximalen Schutz sollte dessen Echtheit nicht nur statisch, sondern auf Basis eines videobasierten Verfahrens und künstlicher Intelligenz geprüft werden. Zusätzlich sollte dies mit einer Prüfung der Sicherheitsmerkmale des NFC-Chips kombiniert und dessen Inhalte ausgelesen werden, um diese mit den optischen Angaben und dem biometrischen Foto abzugleichen. Letzteres dient im weiteren Prozess dazu, die Identität der Person zu verifizieren und potenzielle Deepfakes zu erkennen.

Identitätsüberprüfung mittels Video-Stream: Viele Anbieter bieten die Option, die Identität der Person vor dem Computer beziehungsweise Handy mittels eines Selfies zu prüfen, indem die biometrischen Merkmale des Selfies mit denen des ID-Dokuments abgeglichen werden. Ein solches vereinfachtes Verfahren spielt Betrügern in die Karten, da nur wenig Material vorliegt, um Deepfakes wirksam zu erkennen. Es sollte von Unternehmen besser ein videobasiertes Verfahren verwendet werden, da dies die Möglichkeiten zur Erkennung von Deepfakes deutlich erhöht. Über das gesamte Video kann so, wie oben beschrieben, zum Beispiel der Hintergrund geprüft werden und wie sich das Gesicht und dessen Konturen von ihm abheben. Algorithmen können so zum Beispiel erkennen, dass der Hintergrund ein eingebettetes, statisches Bild ist, was bei Deepfakes häufig der Fall ist. Die Beleuchtung ist ein weiterer guter Ansatzpunkt. So kann man zum Beispiel prüfen, ob ein Schatten im Raum nicht nur das Gesicht, sondern auch andere Objekte sowie den Hintergrund des Bildes, beeinflusst.

Liveness-Test: Ein weiteres, effektives Instrument zum Erkennen von Deepfakes ist ein sogenannter „Liveness"-Test. Hiermit wird sichergestellt, dass die biometrischen Merkmale tatsächlich von einer lebenden Person stammen und nicht von einer künstlichen Person. Üblicherweise wird hierbei die zu identifizierende Person aufgefordert eine zufällige, bestimmte Bewegung zu machen und/oder bestimmte zufällige Wörter zu sprechen.

> „Der Einsatz digitaler Video- und Auto-Ident-Lösungen wird zunehmend unumgänglich."

Aktuell sind quasi keine Deepfake-Systeme in der Lage, ihnen unbekannte, zufällige Bewegungen und Sprache in Echtzeit so realistisch, flüssig und fehlerfrei zu berechnen, dass dies nicht von spezialisierten Prüfalgorithmen als Fälschung erkannt werden könnte.

Deepfakes werden immer verbreiteter und einfacher zu erstellen. Auch wenn sehr gute, realistische Deepfakes mitunter nur sehr aufwändig zu erstellen sind, müssen sich Unternehmen dennoch mit der kontinuierlich steigenden Gefahr auseinandersetzen, die von Identitätsdiebstahl und Online-Betrug ausgeht. Je besser die entsprechenden Algorithmen werden, umso weniger wird es selbst geschulten Mitarbeiterinnen und Mitarbeitern möglich sein, gut gemachte Deepfakes mit bloßem Auge zu erkennen. Der Einsatz digitaler Video- und Auto-Ident-Lösungen, die mittels KI-basierter Authentizitätsprüfung von ID-Dokumenten, dem Abgleich biometrischer Merkmale und „Liveness Detection" effektiven Schutz vor Deepfakes bieten, wird zunehmend unumgänglich.

Christian Gerber

Head of Product, Kerberos Compliance-Managementsysteme GmbH, Köln

Dokumente, Bilder, Tonaufnahmen und Videos werden wohl schon so lange gefälscht, wie sie existieren. Zumindest erforderte das Retuschieren aber vormals Zeit, Präzision und Können in einem Maß, das nur wenige beherrschten. Durch frei verfügbare Technologie und immer weiter steigende Rechenleistung wird es jedoch zunehmend einfacher, mit bloßem Auge – zumindest auf den schnellen Blick – nicht identifizierbare Fälschungen von Bildern, Videos und Tonaufnahmen, sogenannte Deepfakes, zu erstellen oder diese sogar live im Gespräch mit einer weiteren Person zu produzieren. Zwar benötigen perfekte Fälschungen weiterhin Zeit und Aufwand, für Banken können jedoch schon einfache Deepfakes zum Problem werden, wenn sie sich nicht durch entsprechende Prüfverfahren bei Onboarding- und KYC-Prozessen absichern. Der Autor des vorliegenden Beitrags gibt einen Überblick über die Erstellung und Verbreitung von Deepfakes und zeigt einige Mittel und Wege auf, wie man Fälschungen erkennen kann. (Red.)