Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

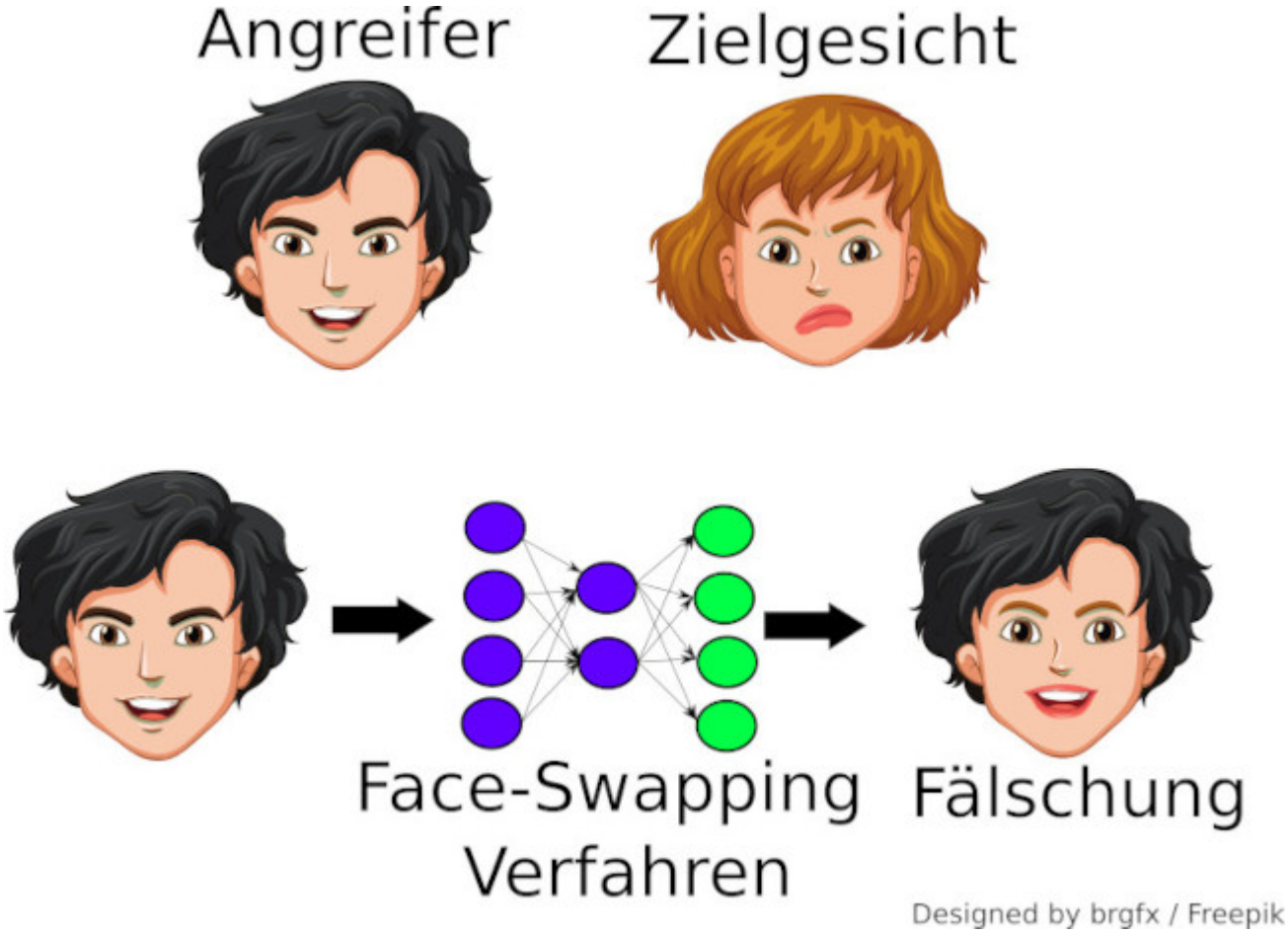
Deepfakes - Gefahren und Gegenmaßnahmen

Verfahren zur Manipulation von medialen Identitäten existieren bereits seit vielen Jahren. So ist es allgemein bekannt, dass Bilder durch vielfältige Methoden manipuliert werden können. Lange Zeit war es sehr aufwändig, dynamische Medien, wie Videos oder Audiomitschnitte qualitativ hochwertig zu manipulieren. Durch Methoden aus dem Bereich der Künstlichen Intelligenz (KI (Künstliche Intelligenz)) ist dies heute jedoch deutlich einfacher und Fälschungen können mit vergleichsweise wenig Aufwand und Expertise in einer hohen Qualität erstellt werden. Aufgrund der Nutzung von tiefen neuronalen Netzen (englisch: deep neural networks), werden solche Verfahren umgangssprachlich als „Deepfakes“ bezeichnet.

Methoden zur Manipulation medialer Identitäten können somit im Wesentlichen in die drei Medienformen Video/ Bild, Audio und Text untergliedert werden. Die folgenden Ausführungen zeigen, welche Angriffsmethoden es nach dem aktuellen Stand der Technik gibt, welche Daten für einen erfolgreichen Angriff benötigt werden und welcher Aufwand für die Erstellung von Fälschungen mithilfe von Deepfake-Verfahren notwendig ist.

Fälschung von Gesichtern

Zur Manipulation von Gesichtern in Videos wurden in den letzten Jahren mehrere KI (Künstliche Intelligenz)-basierte Verfahren entwickelt. Diese verfolgen entweder das Ziel Gesichter in einem Video zu tauschen („Face Swapping“), die Mimik/ Kopfbewegungen einer Person in einem Video nach Wunsch zu kontrollieren („Face Reenactment“), oder neue (Pseudo-)Identitäten zu synthetisieren.



Beim Face-Swapping wird das Gesicht einer Zielperson in das eines Angreifers eingefügt, wobei der Gesichtsausdruck des Angreifers beibehalten werden soll.
Quelle: brgfx / Freepik Zusammenstellung: BSI

Beim „Face Swapping“ Verfahren, dargestellt in der Abbildung oben, besteht das Ziel darin, aus der Eingabe eines Gesichts einer Person, ein Gesichtsbild einer anderen Person mit derselben Mimik, Gesichtsbeleuchtung und Blickrichtung zu erzeugen. Hierfür wird in gängigen öffentlichen Softwarebibliotheken ein Autoencoder-Verfahren als Modell verwendet. Diese neuronalen Netze lernen aus einem Gesichtsbild die relevanten Mimik- und Beleuchtungsinformationen kodiert zu extrahieren und aus den kodierten Informationen ein entsprechendes Gesichtsbild zu erzeugen. Mittlerweile können mit gängigen kommerziellen Grafikkarten Modelle mit hoher Bildauflösung trainiert werden, welche Nahaufnahmen von Gesichtern in FullHD-Videos behandeln können. Diese Modelle sind auch teilweise, und mit einem geringen Zeitversatz, in Echtzeit zum Gesichtstausch einsatzfähig. Als Trainingsmaterial werden dabei nur wenige Videominuten einer Zielperson benötigt. Allerdings müssen diese von hoher Qualität sein und möglichst verschiedene Gesichtsmimiken und Perspektiven enthalten, damit diese vom Modell zur Manipulation gelernt werden können.

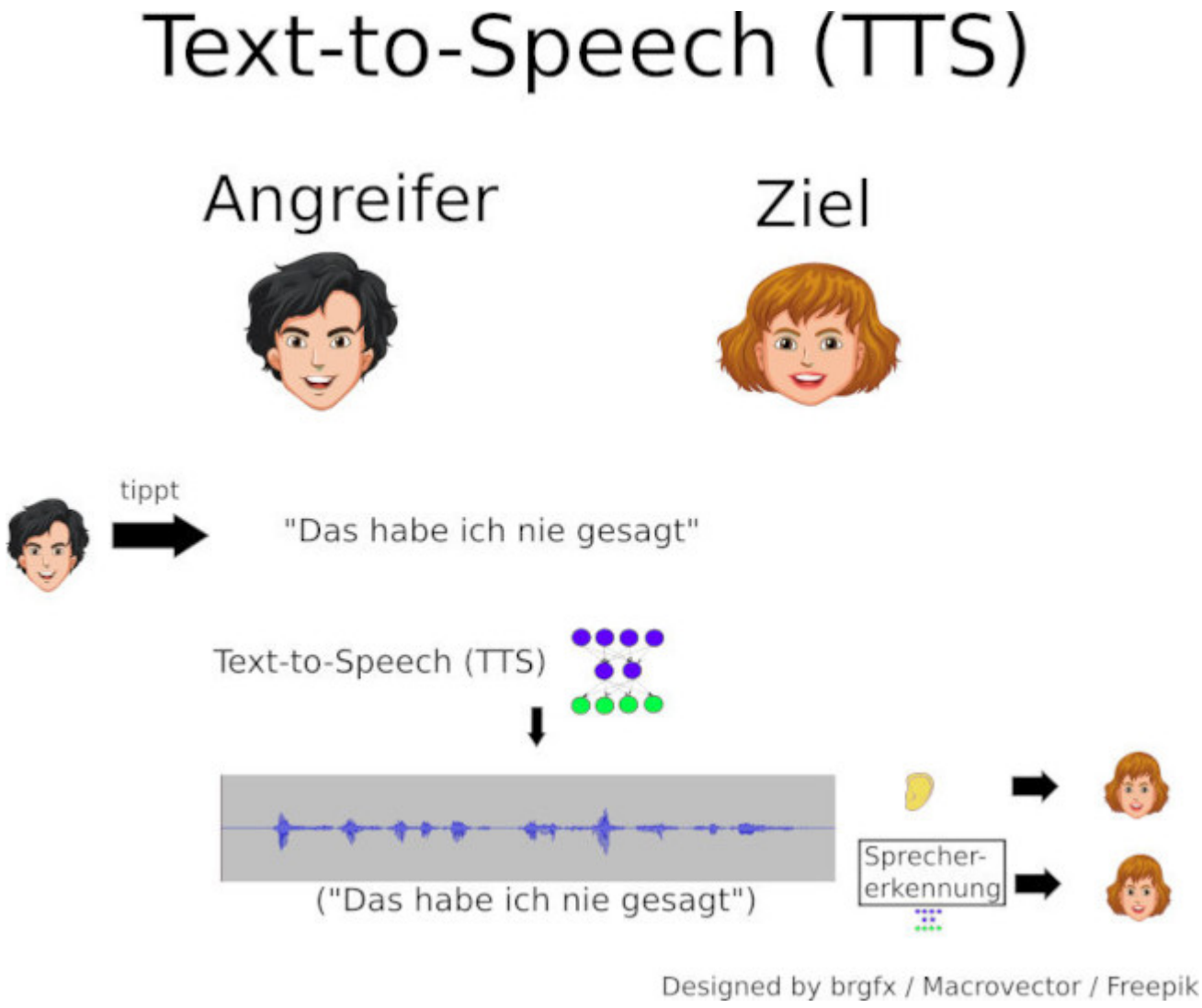
Beim „Face Reenactment“ werden die Kopfbewegung, Mimik oder Lippenbewegung einer Person manipuliert. Dies ermöglicht es, visuell täuschend echte Videos zu erstellen, bei denen eine Person Aussagen trifft, die sie in der Realität nie getätigt hat. Populäre Verfahren erreichen dies durch Erzeugung eines 3D (dreidimensional)-Modells des Gesichts der Zielperson anhand eines Videostreams. Diese der Manipulator dann beliebig mit seinem eigenen Videostream kontrollieren und täuschend echte

Gesichtsausdrücke bei der Zielperson erzeugen.

Bei der Synthetisierung von Gesichtsbildern können neue Personen erzeugt werden, die in der Realität nicht existieren. Gängige Verfahren beschränken sich bisher noch auf einzelne Bilder, welche aber bereits Nahaufnahmen in einer hohen Bildauflösung und Detailtiefe erzeugen können.

Fälschung von Stimmen

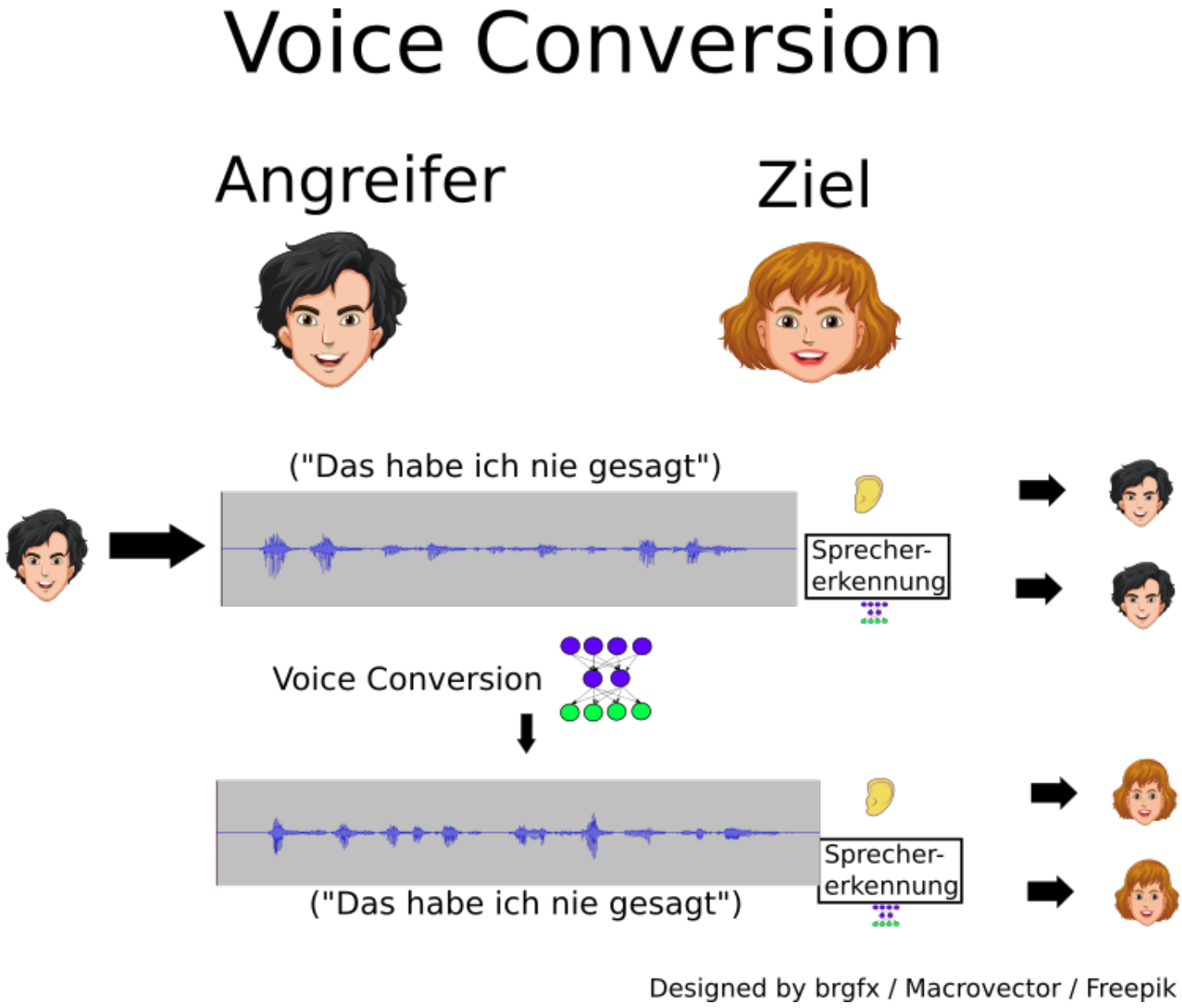
Für die Erstellung von manipulierten Stimmen sind insbesondere „Text-to-Speech (TTS (Text To Speech))“- und „Voice Conversion (VC (Voice Conversion))“-Verfahren von großer Bedeutung.



Beim Text-to-Speech-Verfahren wird zu einem vorgegebenen Text ein Audio-Signal erzeugt, welches sich sowohl für den Menschen als auch für eine automatische Sprechererkennung wie eine Zielperson anhört.
Quelle: brgfx / Macrovector / Freepik Zusammenstellung: BSI

Die prinzipielle Funktionsweise von „Text-to-Speech“-Verfahren ist in der ersten Abbildung skizzie Hierbei kann ein Anwender einen Text vorgeben, welcher durch das TTS (Text To Speech)-System

verarbeitet und in ein Audio-Signal umgewandelt wird. Der semantische Inhalt dieses Signals entspricht dem des vorgegebenen Textes und die sprecherspezifischen Charakteristika entsprechen im Idealfall einer durch den Anwender spezifizierten Zielperson. Hiermit können prinzipiell sowohl Menschen als auch automatisierte Sprecherkennungsverfahren getäuscht werden.



Bei einem Voice Conversion Verfahren wird hingegen ein Audiosignal zu einer Zielstimme konvertiert. Quelle: brgfx / Macrovector / Freepik Zusammenstellung: BSI

Die prinzipielle Funktionsweise eines „Voice-Conversion"-Verfahrens ist in der zweiten Abbildung skizziert. Hierbei hat ein Anwender die Möglichkeit dem VC (Voice Conversion)-System ein Audio-Signal vorzugeben, welches durch dieses zu einem manipulierten Audiosignal konvertiert wird. Dieses erzeugte Audio-Signal hat dabei den gleichen semantischen Inhalt wie das Ursprungssignal, unterscheidet sich jedoch in der zu hörenden Charakteristik des Sprechers/ der Sprecherin. Diese gleicht hierbei im Idealfall einer durch den Angreifer ausgewählten Zielperson.

Damit diese Verfahren funktionieren, müssen sie zunächst mittels Trainingsdaten trainiert werden. der benötigten Daten unterscheidet sich je nach Angriffsart, wobei alle Verfahren die Gemeinsamkei

haben, dass von der Zielperson Audio-Aufnahmen in einer möglichst hohen und konstanten Qualität benötigt werden.

Da sowohl TTS (Text To Speech)-, als auch VC (Voice Conversion)-Verfahren in der Regel durch komplexe neuronale Netze umgesetzt werden, sind für das Training mehrere Stunden Audiomaterial an Trainingsdaten der Zielperson notwendig, um eine hohe Qualität zu erreichen. Allerdings gibt es Möglichkeiten, die von der Zielperson benötigten Daten auf wenige Minuten zu verkürzen, indem große Datenbanken anderer Personen als Hilfsdaten verwendet werden. Moderne Forschungsansätze arbeiten an Verfahren, welche lediglich wenige Sekunden Audiomaterial der Zielperson und keinen erneuten Trainingsprozess benötigen, wobei dies bisher zu Lasten der Qualität der Ausgabe geht.

Fälschung von Texten

Verfahren zur Generierung von Texten, welche auf tiefen neuronalen Netzen basieren, schaffen es durch neue KI (Künstliche Intelligenz)-Modelle, große Textdatenbanken und eine hohe Rechenleistung, lange und zusammenhängende Texte zu schreiben. Bei diesen kann auf den ersten Blick nicht unterschieden werden, ob sie von einem Menschen oder von einer Maschine geschrieben wurden. Meist sind nur wenige einleitende Wörter notwendig, aus denen das Modell eine mögliche, plausible Fortsetzung des Texts generiert. Damit können Nachrichten verfasst, Blog-Einträge erzeugt, oder auch Chat-Antworten generiert werden.

Noch sind die notwendigen Ressourcen zum Training des Systems und Anwendung der leistungsstarken Modelle jenseits dessen, was im Verbraucherbereich üblich ist. Daher müssen Privatpersonen auf öffentlich zugängliche Clouddienste zurückgreifen. Bei fortschreitender Weiterentwicklung der Technik ist damit zu rechnen, dass diese Einsatz in Chatbots oder Social Bots finden, um einen fiktiven Gesprächspartner zu simulieren.

Mögliche Bedrohungsszenarien

Mittels der beschriebenen Verfahren ist es heute auch teilweise für technisch versierte Laien möglich, mediale Identitäten zu manipulieren, wodurch sich zahlreiche Bedrohungsszenarien ergeben:

- **Überwindung biometrischer Systeme:** Da es mittels Deepfake-Verfahren möglich ist, mediale Inhalte mit den Charakteristika einer Zielperson zu erstellen und diese Verfahren teilweise bereits in Echtzeit lauffähig sind, stellen sie eine hohe Gefahr für biometrische Systeme dar. Insbesondere bei Fernidentifikationsverfahren (z.B. (zum Beispiel) der Sprechererkennung über das Telefon oder der Videoidentifikation) scheinen solche Angriffe erfolgversprechend, da ein potentieller Verteidiger lediglich das Ausgangssignal erhält. Jedoch hat er keine Kontrolle über die Aufnahmesensorik oder die am aufgenommenen Material durchgeführten Änderungen.

- **Social Engineering:** Deepfake-Verfahren können außerdem dazu verwendet werden, gezielte Phishing-Angriffe („Spear-Phishing“) durchzuführen, um Informationen und Daten zu gewinnen.
  Auch kann ein Angreifer diese Technologie zur Durchführung von Betrug und zur Abschöpfung finanzieller Mittel nutzen. Beispielsweise könnte er eine Person mit der Stimme von deren Führungskraft anrufen, um eine Geldtransaktion auszulösen („CEO (Chief Executive Officer)-Fraud“).
- **Desinformationskampagnen:** Mittels Deepfake-Verfahren ist es potentiell möglich, glaubwürdige Desinformationskampagnen durchzuführen, indem manipulierte Medieninhalte von Schlüsselpersonen erzeugt und verbreitet werden.
- **Verleumdung:** Durch die Möglichkeit Medieninhalte zu generieren, die Personen beliebige Aussagen treffen lassen und sie in beliebigen Situationen darstellen, kann der Ruf einer Person durch die Verbreitung von Unwahrheiten nachhaltig geschädigt werden.

Beispiel-Video

Das folgende Video zeigt beispielhaft drei verschiedene Fälschungsmethoden.

Zum einen wird in dem Video ein Face-Swapping-Verfahren angewendet, um das Gesicht eines An

mit Arne Schönbohm zu tauschen, zum Produktionszeitpunkt des Videos Präsident des BSI (Bundesamt für Sicherheit in der Informationstechnik). Dazu wurden im Vorfeld ca. (ungefähr) 5-10 Minuten lange Videoaufnahmen von beiden Personen gemacht, welche für das Training des KI (Künstliche Intelligenz)-Modells verwendet wurden. Insbesondere zeigt das Video, dass es schon heute möglich ist, eine solche Fälschung in Echtzeit mit einer vergleichsweise hohen Qualität zu erstellen.

Zum anderen sind in dem Video zwei Audio-Fälschungen enthalten. Bei einer dieser Fälschungen wurde ein Text-to-Speech-Verfahren verwendet, um Audio-Segmente der zu fälschenden Stimme zu erstellen. Außerdem wurde die Stimme des „Off-Sprechers“ mittels eines Voice-Conversion-Verfahrens in die Stimme des damaligen BSI (Bundesamt für Sicherheit in der Informationstechnik)-Präsidenten konvertiert. Für das Training dieser Systeme wurden ca. (ungefähr) 10 Minuten Audio-Material des Originals verwendet, das aus öffentlichen Videos extrahiert wurde und nur von mittlerer Qualität war.

In den folgenden Audio-Segmenten ist zunächst die Original-Aufnahme eines Sprechers zu hören. Im darauffolgenden Segment ist ein mittels eines TTS (Text To Speech)-Verfahrens generiertes Audio-Signal dieses Sprechers zu hören. Abschließend enthält das letzte Audio-Segment eine mittels eines VC (Voice Conversion)-Verfahrens manipulierte Version der Original-Aufnahme, welche Arne Schönbohm als Zielsprecher hat.

- Das Original:

- TTS-Version der Stimmenmanipulation:

- Arne Schönbohm-konvertierte Version der Sprachausgabe:

In den Audiodateien gesprochener Text: „Schönen guten Tag, meine Damen und Herren. Ich bin nicht echt, noch können Sie das wahrnehmen. Aber mit zunehmender Reife der Technologie, wird Ihnen das sehr, sehr schwer fallen."

Gegenmaßnahmen

Es gibt viele Ansätze, um sich gegen die beschriebenen Methoden zu verteidigen, wobei sich diese in die zwei Kategorien Prävention und Detektion untergliedern lassen.

1. Prävention

Gegenmaßnahmen aus dem Gebiet der Prävention zielen darauf ab, das Risiko eines erfolgreichen Angriffs mittels Deepfakes zu senken:

Aufklärung

Eine zentrale Maßnahme gegen Deepfake-Angriffe stellt die Schulung potentiell betroffener Personen dar. Zum einen ist davon auszugehen, dass das Wissen über die Möglichkeit eines solchen Angriffs eine differenzierte Einschätzung der Echtheit des gesehenen oder gehörten Materials unter Berücksichtigung der Quelle ermöglicht. Zum anderen erzeugen viele Deepfake-Verfahren teilweise deutliche Artefakte. Durch die Kenntnis dieser Artefakte kann die Erkennung von Fälschungen signifikant gesteigert werden. Insbesondere bei Echtzeitanwendungen hat ein Angreifer nicht die Möglichkeit, mit Artefakten behaftetes Material manuell zu bereinigen.

Typische Artefakte bei Gesichtsmanipulationen

- **Sichtbare Übergänge:** Bei einem Face-Swapping-Verfahren wird ein Gesicht der Zielperson in den Kopf einer anderen Person eingesetzt. Dadurch kann es zu sichtbaren Artefakten an der Naht rund um das Gesichts kommen. Ebenso ist es möglich, dass die Hautfarbe und -textur an diesem Übergang wechselt oder dass sich teilweise das Ursprungsgesicht in manchen Frames am Gesichtsrand durch doppelte Augenbrauen bemerkbar macht.
- **Scharfe Konturen verwaschen:** Häufig kommt es noch vor, dass Face-Swapping-Verfahren nicht richtig lernen, scharfe Konturen, wie sie in den Zähnen oder im Auge vorkommen, zu erzeugen. Bei genauem Hinsehen wirken diese auffällig verwaschen.
- **Begrenzte Mimik, unstimmige Belichtung:** Auf Grund einer beschränkten Datenlage kann es dazu kommen, dass ein Modell nur beschränkt fähig ist manche Gesichtsausdrücke oder Beleuchtungssituationen korrekt darzustellen. Häufig ist die Profilansicht eines Gesichts unzureichend erlernt, sodass ein starkes Drehen des Kopfes zu Bildfehlern führen kann, bei welchen zum Beispiel das Gesicht verwaschener wird.

Typische Artefakte bei synthetischen Stimmen

- **Metallischer Sound:** Zahlreiche Verfahren, erzeugen ein Audio-Signal, das vom menschlichen Gehör als „metallisch“ wahrgenommen wird.
- **Falsche Aussprache:** Häufig können TTS (Text To Speech)-Verfahren nicht alle Wörter korrekt aussprechen. Dies kann beispielsweise passieren, wenn ein TTS (Text To Speech)-Verfahren für die deutsche Sprache trainiert wurde, aber ein englisches Wort ausgesprochen werden soll.
- **Monotone Sprachausgabe:** Insbesondere wenn die Trainingsdaten für ein TTS (Text To Speech)-System nicht ideal sind, kann das erzeugte Audio-Signal sehr monoton hinsichtlich der Betonung der Wörter sein.
- **Falsche Sprechweise:** Meist sind Fälschungsverfahren vergleichsweise gut dafür geeignet, die Klangfarbe einer Stimme zu fälschen, haben jedoch häufig Probleme damit, die spezifischen Charakteristika der Stimme zu fälschen, sodass beispielsweise Akzente oder Betonungen von Wörtern nicht denen des Zielsprechers/ der Zielsprecherin entsprechen.
- **Unnatürliche Geräusche:** Sofern ein Fälschungsverfahren Eingangsdaten erhält, die stark von den beim Training verwendeten abweichen, kann das Verfahren unnatürliche Geräusche erzeugen. Dies kann beispielsweise ein zu langer Text bei einem Text-to-Speech-Verfahren oder Stille bei einem Voice-Conversion-Verfahren sein.
- **Hohe Verzögerung:** Die meisten Verfahren zur Erzeugung von synthetischen Stimmen müssen zunächst einen Teil des zu erzeugenden semantischen Inhalts als Eingangsdaten empfangen, um ein qualitativ hochwertiges Ergebnis zu erzeugen. Dies führt dazu, dass qualitativ hochwertige Fälschungen in vielen Fällen mit einer gewissen zeitlichen Verzögerung einhergehen, da dieser semantische Inhalt zunächst ausgesprochen und erfasst werden muss, bevor er von einem VC (Voice Conversion)/ TTS (Text To Speech) Verfahren verarbeitet werden kann.
- Um die Fähigkeit, manipulierte Audio-Daten zu erkennen, zu trainieren, kann beispielsweise die von **Fraunhofer AISEC** entwickelte Anwendung verwendet werden.

Kryptographie

Kryptographische Verfahren bieten die Möglichkeit, die Quelle des Materials eindeutig an eine Identität zu binden. Dies ermöglicht die sichere Zuordnung zu einer (vertrauenswürdigen) Quelle (Authentizität) und stellt sicher, dass Manipulationen des Materials nach der Absicherung sofort auffallen (Integritätsschutz). Hierdurch kann jedoch nicht verhindert werden, dass die Quelle selbst das Material zuvor manipuliert. Aktuelle Entwicklungen beschäftigen sich beispielsweise mit der Erstellung einer digitalen Signatur beim Aufnahmeprozess, wodurch sichergestellt wird, dass das Material nicht mehr nach der Aufnahme manipuliert wurde.

Gesetzlich

Gesetzliche Regelungen können eine Hürde darstellen, Deepfakes ungekennzeichnet in Umlauf zu bringen. Insbesondere wird in dem Regulierungsentwurf der EU (Europäische Union)-Kommission zu KI (Künstliche Intelligenz)-Systemen gefordert, dass alle mit der Deepfake-Technologie erstellten Materialien als solche gekennzeichnet werden müssen.

2. Detektion

Gegenmaßnahmen aus dem Bereich der Detektion zielen darauf ab, mittels Deepfake-Verfahren manipulierte Daten als solche zu erkennen.

Medienforensisch

Mittels Methoden aus der Medienforensik ist es möglich, Artefakte zu detektieren, welche bei der Verwendung von Manipulationsmethoden auftreten. Hiermit ist es für Expertinnen und Experten möglich, Fälschungen nachvollziehbar zu erkennen.

Automatisierte Detektion

In der Forschungsliteratur wurden in den letzten Jahren zahlreiche Methoden zur automatisierten Detektion von manipulierten Daten veröffentlicht. Diese Verfahren basieren in der Regel auf Techniken aus dem Gebiet der künstlichen Intelligenz, insbesondere den tiefen neuronalen Netzen. Aufgrund dessen müssen diese Verfahren anhand großer Datenmengen trainiert werden. Nach der Trainingsphase kann das Modell dazu verwendet werden, für ein Datenbeispiel (zum Beispiel ein Video) zu klassifizieren, ob dieses manipuliert wurde oder nicht.

Herausforderungen der (automatisierten) Gegenmaßnahmen

Ein Problem der Gegenmaßnahmen besteht darin, dass diese entweder nicht in allen Situationen angewendet werden können und in der Regel keinen vollständigen Schutz bieten.

Insbesondere bei der Klasse der automatisierten Detektionsmethoden sollte darauf hingewiesen we[illegible],

dass sie häufig nur unter gewissen Rahmenbedingungen zuverlässig funktionieren. Da diese Verfahren in der Regel auf Verfahren der künstlichen Intelligenz basieren, gehen diese Methoden auch mit deren grundsätzlichen Problemen einher:

- **Generalisierbarkeit:** Ein zentrales Problem der meisten Detektionsmethoden ist ihre mangelhafte Generalisierbarkeit. Da die Methoden auf bestimmten Daten trainiert wurden, funktionieren sie häufig auf ähnlichen Daten relativ zuverlässig. Werden jedoch einzelne Parameter verändert, so ist die Korrektheit der Ausgaben häufig nicht gegeben. Ein wichtiges Beispiel eines solchen Parameters, kann der Wechsel zu einer anderen Angriffsmethode, welche nicht in den Trainingsdaten vorhanden war, sein. Dieses Verhalten konnte beispielsweise in der **Deepfake Detection Challenge** (2020) beobachtet werden, in welcher selbst das beste Modell lediglich eine durchschnittliche Genauigkeit von 65,18 Prozent erreichen konnte, wobei eine Genauigkeit von 50 Prozent durch bloßes Raten erreicht werden würde.
- **KI (Künstliche Intelligenz)-spezifische Angriffe:** Ein weiteres zentrales Problem dieser Verfahren besteht darin, dass sie durch KI (Künstliche Intelligenz)-spezifische Angriffe überwunden werden können, wobei insbesondere adversariale Angriffe eine besondere Bedrohung darstellen. Weitere Informationen hierzu: **Sicherer, robuster und nachvollziehbarer Einsatz von KI**
  So kann ein adaptiver Angreifer beispielsweise ein gezieltes Rauschen erstellen, welches über das mittels eines Face-Swapping-Verfahrens manipulierte Bild gelegt wird. Dieses Rauschen kann so klein sein, dass es für den menschlichen Betrachter nicht zu bemerken ist, hat jedoch für das Detektionsverfahren den Effekt, dass es die Fälschung nicht als solche klassifiziert. Solche Angriffe lassen sich nicht komplett vermeiden, jedoch sollten sie bei der Erstellung von Detektionsverfahren berücksichtigt und die Hürde für einen Angreifer erhöht werden.

Ausblick

Die Technologie zur Fälschung medialer Identitäten hat sich in den letzten Jahren insbesondere durch die Fortschritte im Bereich der künstlichen Intelligenz deutlich weiterentwickelt. Aktuelle Forschungsergebnisse deuten darauf hin, dass sich dieser Trend weiter fortsetzen wird, sodass die manuelle Erkennung von Fälschungen in Zukunft immer schwieriger werden wird. Auch ist davon auszugehen, dass sich die Menge der von der angegriffenen Person benötigten Daten stetig verringern wird. Für technisch versierte Laien ist es bereits heute möglich, qualitativ hochwertige Fälschungen zu erstellen. Es ist jedoch davon auszugehen, dass sich die benötigte Expertise und der notwendige Aufwand zur Erstellung von Fälschungen durch die Verbesserung und erhöhte Verfügbarkeit an öffentlichen Tools stetig verringern wird, sodass sich die Häufigkeit von Angriffen mittels dieser Technologie signifikant erhöhen könnte. Aus diesen Gründen ist es von hoher Bedeutung, dass die aufgeführten Gegenmaßnahmen weiterentwickelt und in Kombination nach einer applikationsspezifischen Auswahl eingesetzt werden.

Ähnliche Themen

Kriterienkatalog für KI-Cloud-Dienste – AIC4

Biometrie als KI-Anwendungsfeld

Zurück zu Künstliche Intelligenz

TOP