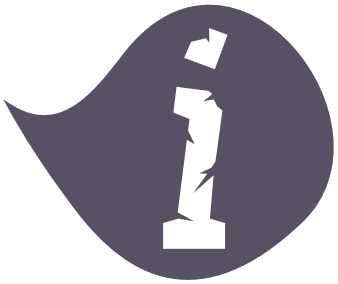
Täuschend echt: Deepfakes im Fokus



Millionen Menschen sahen in sozialen Medien eine Anzeige, in der Taylor Swift für ein Gewinnspiel warb, bei dem man angeblich teure Designer-Kochtöpfe gewinnen konnte. Die Anzeige, Taylor Swift, sogar ihre Stimme wirkten dabei täuschend echt. Es handelte sich aber um eine Fälschung. Betrügerinnen und Betrüger nutzten eine Künstliche Intelligenz, um ein Bild von Taylor Swift zu erzeugen und ihre Stimme nachzuahmen, ein sogenanntes Deepfake. Die Fake-Promi-Werbung war eine Abofalle. Der Fake wurde über 40 Millionen Mal angesehen und geteilt, bevor er entdeckt und gelöscht wurde. Taylor Swift wusste nicht, dass man ihren Körper und ihre Stimme benutzt hatte. Dies ist nur ein Beispiel für den Missbrauch der Deepfake-Technologie, die in den falschen Händen und mit böswilligen Absichten enormen Schaden anrichten kann.

Was sind Deepfakes – und wie werden sie erzeugt?

Deepfakes sind Bilder, Videos oder Audiodateien, die mit Hilfe von *Künstlicher Intelligenz* (KI) erstellt werden. Der Begriff setzt sich zusammen aus „Deep Learning“[1] (‚tiefes, umfassendes Lernen‘) und „Fake“, an dieser Stelle eine KI-Technik, die reale Personen, Stimmen oder Situationen fälschen kann. Deepfakes sind in der Lage, die Realität täuschend echt nachzuahmen.

Dazu werden beispielsweise Fotos und Videos eines Promis in das Programm eingegeben. Das Programm „merkt“ sich die verschiedenen Gesichtszüge, biometrischen Daten (einzigartige körperliche Merkmale), Stimmmuster und Bewegungen der Person. Es erkennt Zusammenhänge und Muster. Man kann sich das so vorstellen: Wenn der Mund lächelt, sind auch die Augenbrauen hochgezogen, wenn im Liedtext ein „O“ vorkommt, formt sich der Mund entsprechend. Je mehr Beispieldaten für das Training des Programms benutzt werden, desto genauer kann die KI den Promi analysieren und später nachbilden. Den zweiten Schritt kann man sich etwa so vorstellen: Das Programm „zerlegt“ den Star in Einzelteile: Augen in verschiedenen Blickrichtungen, Mundwinkel mal lächelnd, mal ernst, Augenbrauenhaltungen, Körperformen, Posen und so weiter. In einem dritten Schritt setzt die KI den Promi neu zusammen und kann ihn je nach Anweisung Songtexte singen und Dinge tun lassen, die so niemals passiert sind. Die KI ist ebenfalls in der Lage, mehrere Datensätze zu verbinden. So kann zum Beispiel das Gesicht von Selena Gomez auf den Körper von Donald Trump montiert werden.

Formen von Deepfakes

Man unterscheidet, je nachdem, was gefälscht wird, zwischen verschiedenen Formen von Deepfakes:

[1] Eine mathematische Methode, die aus großen Datenmengen lernt und darin Muster erkennt, um Entscheidungen ähnlich wie ein menschliches Gehirn zu treffen. Deep Learning Methoden werden häufig zur Gesichts-, Objekt- oder Spracherkennung eingesetzt.

- **Gesichtsübernahme (Face swapping, Face replacement)**
  Hierbei wird das Gesicht von Person 1 in das Gesicht von Person 2 eingefügt. Der Gesichtsausdruck von Person 2 bleibt dabei erhalten. Die zweite Person agiert also mit dem Gesicht der ersten Person. Auf diese Weise können Echtzeitvideos erstellt werden, in denen man z. B. in einer Videokonferenz mit dem Gesicht einer anderen Person etwas sagt.

- **Gesichtsverfälschung (Face reenactment)**
  Dabei werden Mimik, Lippenbewegungen oder Kopfbewegungen einer Person verändert, manipuliert und neu berechnet. Durch Gesichtsfälschung oder Gesichtsübernahme ist es möglich, visuell täuschend echte Videos zu erstellen und beispielsweise Personen Aussagen in den Mund zu legen, die sie nie gemacht haben.

- **Gesichtserfindung (Face generation)**:
  Die Software lernt aus Millionen von Gesichtern, wie ein Mensch aussieht und kann aus den gelernten Informationen Personen zusammensetzen, die in der Realität nicht existieren. Die Ergebnisse sind von echten Fotos fast nicht mehr zu unterscheiden.

- **künstliche Nacktbilder (Deep nudes)**
  Auf realen Fotografien retuschiert die KI die Kleidung der abgebildeten Person und ersetzt sie durch nackte Haut. So können Nacktbilder und -videos von Personen gefälscht werden.

- **Stimmenfälschung (Speech synthesis)**
  Bei diesem Verfahren wird ein Text („Text-to-Speech“) oder ein Audiosignal („Voice-Conversion“) vorgegeben. Die KI wandelt dann den Text oder das Audio in ein Audiosignal um, das wie die Stimme der Zielperson klingt. So kann man beispielsweise mit der Stimme einer anderen Person telefonieren.

Wo liegen die Gefahren von Deepfakes?

KI-generierte Bilder oder Videos können eingesetzt werden, um Desinformation und politischer Propaganda mehr Glaubwürdigkeit zu verleihen.

1) Deepfakes können falsche Bilder erzeugen und die Realität verzerren

Im Internet kursieren zahlreiche Deepfakes, die Politikerinnen und Politiker in skurrilen Situationen darstellen, z. B. Donald Trump und Vladimir Putin, die gemeinsam ein Glas Wein trinken oder Bilder, die Donald Trumps Verhaftung zeigen. Solche Bildbotschaften sprechen unsere Emotionen an und wollen unsere politische Meinung manipulieren. Tatsächlich sind sie aber gefälscht und verzerren die Realität.

2) Personen können falsche Worte in den Mund gelegt werden

In einem Deepfake-Video sagt Bundeswirtschaftsminister Robert Habeck in einer Talkshow zur Moderatorin Sandra Maischberger sie solle „endlich mal ihre dumme Fresse halten“. Habeck wird in dieser Fälschung in einem extrem negativen Licht gezeigt.

3) Personen des öffentlichen Lebens können echtes Material als Fälschung abtun

Immer wieder gibt es Aufnahmen von Politikerinnen und Politikern, die etwas tun oder sagen, was nicht gut ankommt. Um sich aus der Affäre zu ziehen, könnten sie behaupten, es handele sich um ein Deepfake und der Vorfall habe nie stattgefunden. Hier wird deutlich, dass allein

die Existenz der Deepfake-Technologie eine Quelle für Desinformation sein kann, selbst wenn sie nicht eingesetzt wird.

Cybermobbing
Im schulischen und privaten Kontext heben Deepfakes die Möglichkeiten des Cybermobbings auf ein neues Level. Sie machen es möglich, Mitschülerinnen und Mitschüler sowie Lehrkräfte in allen erdenklichen Kontexten täuschend echt darzustellen, bloßzustellen und zu demütigen, beispielsweise indem Fake-Nacktbilder erstellt werden. Da heute alle Deepfake-Material mit Smartphone-Apps erzeugen können, betreffen Deepfakes zunehmend Privatpersonen.

Kann die Technologie hinter Deepfakes auch positives bewirken?

Natürlich – und deshalb ist sie nicht verboten. Man kann zum Beispiel KI-generierte historische Persönlichkeiten wie Julius Cäsar oder Marie Curie zum Leben erwecken oder Schauspielerinnen und Schauspieler in Filmen und Serien für Rückblenden verjüngen. Beim Online-Shopping können Kleidungsstücke mit Hilfe von Ganzkörper-Modellen anprobiert werden. Im medizinischen Bereich könnten Menschen, die nicht mehr sprechen können, durch Sprachsynthese ihre Stimme zurückerhalten. Es ist auch kein Problem, im Freundeskreis Körper und Gesichter zu vertauschen – solange alle damit einverstanden sind.

Die Rechtslage – wann sind Deepfakes verboten?

Einwilligung ist hier das Stichwort. Generell gilt: Menschen haben ein Recht am eigenen Bild. Das bedeutet: Sobald andere, reale Menschen beteiligt sind, muss man um Erlaubnis fragen, wenn man diese Menschen in ein Deepfake einbezieht (vgl. StGB § 201a). Ebenso ist es verboten, jemandem etwas zu unterstellen, was er nicht getan oder gesagt hat (vgl. StGB §§ 185, 186, 187). Fake-Pornos können als sexueller Missbrauch eingestuft werden, wobei pornografische Darstellungen von Minderjährigen grundsätzlich verboten sind (vgl. StGB § 184a-c). Noch haben die gesetzlichen Bestimmungen rund um Deepfakes Lücken und müssen noch an diese neue Technologie angepasst werden.

Wie erkennt man Deepfakes?

Es wird immer schwieriger, Deepfakes zu identifizieren, da sie immer realistischer erscheinen. Es gibt jedoch typische Merkmale und Hilfen, wie man Deepfakes auf die Spur kommen kann.

Beim Face-Swapping kommt es zu sichtbaren Übergängen und Fehlern, sogenannten Artefakten, rund um den Kopf der Ursprungs- und Zielperson: Die Haar- oder Hautfarbe wechselt, der Übergang von Gesicht und Haaren ist nicht stimmig, das Ursprungsgesicht scheint durch oder die Gesichtsproportionen passen nicht zueinander. Auch bei der Erzeugung von synthetischen Stimmen entstehen Fehler: Der Klang ist metallisch, die Aussprache ist monoton oder die Betonung ist falsch. Zudem erzeugt die Software oft Hintergrundgeräusche, die nicht zur Situation passen und reagiert bei Echtzeit-Antworten merkwürdig verzögert.

Beim Erkennen von Deepfakes können klassische Strategien für das Überprüfen von Nachrichten eingesetzt werden:

- Logik: Passt das, was man sieht zu dem gewohnten Verhalten der Person?
- Vergleich: Berichten auch seröse Nachrichtenseiten oder Faktencheck-Portale über die Meldung?

- Quellensuche: Kann man herausfinden, wo das Foto/Video zum ersten mal aufgetaucht ist? Hier kann eine Bilderrückwärtssuche hilfreich sein.

Was kann man gegen rechtsverletzende Deepfakes tun?

So können Betroffene handeln
Betroffene können Online-Plattformen auffordern, Deepfakes zu entfernen. Die Plattformen sind nach dem Telemediengesetz und dem Netzwerkdurchsetzungsgesetz dazu verpflichtet, rechtsverletzende Inhalte zu prüfen und gegebenenfalls zu löschen.

Das können Zeuginnen und Zeugen tun
Wer den Verdacht hat, auf einen rechtsverletzenden Deepfake gestoßen zu sein, sollte das Material auf keinen Fall weiterleiten, auch nicht, um andere zu warnen. Denn so erhält der Fake nur noch eine höhere Reichweite. Am Ende bleibt das Bild im Gedächtnis, auch wenn man weiß, dass es nicht echt ist. Damit unterstützt man die Täterinnen und Täter, vergrößert das Leid der Betroffenen und macht sich gegebenenfalls selbst strafbar. Stattdessen kann man das Material bei der Online-Plattform melden oder an Faktencheck-Organisationen wie den Correctiv-Faktencheck, den Faktencheck der dpa den Faktencheck der Deutschen Welle oder Mimikama zur Überprüfung schicken.

Kursieren Deepfakes im privaten Umfeld, kann man sich, ähnlich wie bei (Cyber-)Mobbing, Hilfe holen. Man kann sich an Vertrauenspersonen oder Beratungsstellen wie die NummergegenKummer, Juuuport, HateAid, Jugendschutz.net oder lokale Beratungsangebote wenden. Bei einigen Meldestellen (z. B. HessenGegenHetze) kann man die Beweismaterialien auch direkt hochladen. Um die Beweise zu sichern, sollte man rechtssichere Screenshots anfertigen, die beispielsweise Datum und Uhrzeit beinhalten. Unterstützung bieten auch Seiten im Internet, die über Cybermobbing informieren wie Klicksafe.

Die Macht der Bilder

Mit fortschreitender Verbesserung von Technologien wird es immer schwieriger werden, Deepfakes von der Realität zu unterscheiden. Dabei ist nicht jedes KI-generierte Bild eine Desinformation. Es kommt auf die Absicht an. Gefährlich werden Deepfakes, wenn sie unsere Wahrnehmung der Realität verzerren und manipulieren wollen. Menschen neigen dazu, Bildern mehr zu vertrauen als Texten, sie intensiver wahrzunehmen und stärker mit Emotionen zu verbinden. „Ein Bild sagt mehr als tausend Worte“, sagt ein Sprichwort. Bilder bleiben auch länger im Gedächtnis haften, sie wirken nach und können uns auch dann noch beeinflussen, wenn wir wissen, dass sie Fake sind – einfach, weil sie in unserem Kopf sind. Das macht Deepfakes so gefährlich – und daher müssen wir uns umgewöhnen. Im Journalismus galt das Pressefoto lange Zeit als glaubwürdiges Abbild der Wirklichkeit mit Beweiskraft. Heute müssen wir lernen, die audiovisuellen Medien kritischer zu betrachten. Deepfakes haben das Potenzial, gesellschaftliche Debatten bis hin zu Wahlen zu beeinflussen und die öffentliche Meinung zu manipulieren – vor allem in sozialen Medien, in denen solche Inhalte schnell viral gehen können.

Je mehr Menschen darüber informiert sind, wie Deepfakes funktionieren und wie sie für Desinformation und Cybermobbing genutzt werden können, desto weniger Macht haben Deepfakes über unser Leben. Aufklärungsarbeit über diese relativ neue Technologie steht daher im Mittelpunkt, um den Missbrauch von Deepfakes zu bekämpfen.