Hochschule der Medien SS 2020

Whitepaper zum Thema “Deepfakes”

EDV-NR: 253504A AKTUELLE THEMEN

ARON KÖCHER, MATRIKEL NR: 38372

Einleitung

Bei einer Bitkom Research-Umfrage im Jahr 2018 wurden über 1000 Teilnehmer gefragt, ob sie eher Chancen oder Risiken in der künstlichen Intelligenz sehen. 62% antworteten, dass sie eine klare Chance sehen und fast jeder 5. glaubt, künstliche Intelligenz verändert bereits jetzt unsere Gesellschaft spürbar. Allerdings haben 35% gemeint, dass künstliche Intelligenz ein Risiko darstellen kann [1]. Ein Beispiel für ein potentielles Risiko von künstlicher Intelligenz sind Deepfakes. In der folgenden Ausarbeitung soll sowohl gezeigt werden, was Deepfakes sind, wie sie funktionieren und welche Einsatzgebiete es gibt. Außerdem soll die Frage beantwortet werden, ob und welche Risiken von Deepfakes ausgehen können und wie man die Risiken einschränken kann.

Definition

Als „Deepfake“ werden realistisch wirkende Bild-, Audio oder Videomaterialien bezeichnet, die mit Hilfe von künstlicher Intelligenz manipuliert wurden. Dabei kommen in erster Linie Verfahren aus dem Machine Learning zum Einsatz, speziell dem Deep Learning. Dadurch leitet sich auch der Name ab: Deepfake ist eine Kombination der englischen Wörter „Deep Learning“ und „Fake“ und wurde erstmals 2017 geprägt [2].

Geschichte

Die folgenden Meilensteine zeigen einschlägige Ereignisse im Bereich der Deepfakes: Bereits im Jahr 1997 wurde ein wissenschaftliches Paper zum Thema „Video Rewrite“ veröffentlicht, welches Videomaterial einer sprechenden Person modifiziert, in dem es die Person beim Sprechen einer anderen Tonspur zeigt. Grundlage zur vollständigen automatischen Animation der Gesichter war hierbei maschinelles Lernen [3]. 2016 wurde ein weiteres Paper veröffentlicht, welches es ermöglicht, in Echtzeit die Mimiken eines Gesichts auf eine Person in einem Video zu animieren [4]. Ein Jahr darauf folgte dann die Veröffentlichung eines Programms, bei dem der ehemalige Präsident der USA Barack Obama per Deepfake dargestellt wurde. Hierbei ist deutlich ein Qualitätssprung in der Entwicklung zu erkennen und erste Medien fingen an Deepfakes genauer zu betrachten, um auch auf mögliche Gefahren hinzuweisen [5], [6]. Folglich gab es auch die ersten Deepfakes im privaten Bereich. Ein Nutzer namens „deepfakes“ hat im Dezember 2017 auf der Social-News-Plattform Reddit diverse, eigens erstellte Pornovideos, hochgeladen. In diesen Videos wurden Gesichter von Pornostars mit denen von Scarlett Johansson, Taylor Swift und der Wonder Woman Darstellerin Gal Gadot ausgetauscht [7], [8]. Anfang 2018 folgte daraufhin mit FakeApp eine Desktop Anwendung, die Deepfakes relativ benutzerfreundlich Privatpersonen zugänglich machte [9]. Hierbei wird das Gesicht in einem Video ausgetauscht. Kurze Zeit später folgte dann auch die Bereitstellung von DeepFaceLab, einer Anwendung, die nach eigenen Angaben mehr als 95% des im Internet veröffentlichten Deepfake Materials zu verantworten hat [10]. 2019 wurde die Technologie nochmals weiterentwickelt. In einer Kooperation von Samsung mit den Wissenschaftlern des Skolkovo Institute of Science and Technology ist es ihnen gelungen, eine Verfahrensweise zu entwickeln Deepfakes auf Grundlage eines einzigen Eingabebildes zu erstellen. Während dies bei einem Test mit der „Mona Lisa“ relativ realistisch aussieht, ist es allerdings bei anderem Testmaterial nicht ganz so realistisch. Dies kann allerdings mit deutlich weniger Eingabematerial, als bei früheren Technikständen, verbessert werden [11].

Funktionsweise
Wie in nahezu allen Modellen der künstlichen Intelligenz wird zunächst Ausgangsmaterial benötigt. Dabei bieten sich Bewegtbilder, also Videomaterial von Personen an, die man manipulieren möchte. Je mehr Ausgangsmaterial vorhanden ist, umso besser kann das Training des Modells erfolgen. Ein Videomaterial aus verschiedenen Perspektiven sowie mit verschiedenen Mimiken macht das Ergebnis noch besser.

Es gibt verschiedene Ansätze, um einen Deepfake umzusetzen. Eine gängige Variante ist der Autoencoder. Dieses neuronale Netz erstellt aus einem Eingabebild ein digitales Abbild (auch Encoding genannt). Dabei wird das Eingabebild z.B. durch ein Verfahren wie Max-Pooling verkleinert und auf wesentliche Kernmerkmale des Bildes reduziert. Auf Grundlage des encodierten Bildes wird im Anschluss versucht das Bild möglichst echt künstlich nachzubilden (auch Decoding genannt). Folglich hat das Decoding das Ziel ein identisches, künstliches Abbild des encodierten Eingabebildes zu erzeugen. Um den Autoencoder noch robuster zu machen, ist es außerdem ratsam die Eingabebilder zu verzerren, Auflösungen zu variieren und die Größe zu verändern. Über die Zeit des Trainings wird so das Netz immer besser und der Unterschied zwischen Eingangsbild und Ausgangsbild ist nahezu nicht mehr vorhanden.

Beim Tausch von zwei Personen wird jeweils das Autoencoder Verfahren (bestehend aus Encoder und Decoder) auf beide Personen angewendet. Hierfür wird allerdings der gleiche Encoder verwendet, aber zwei verschiedene Decoder. Während dem Training wird zuerst Bildmaterial von Person A encodiert und im Anschluss mit Decoder A nachgestellt. Anschließend folgt das gleiche Training mit Person B und Decoder B. Nach dem Training wird der Decoder A aus dem Modell von Person A mit dem Decoder von Person B ausgetauscht und man erhält so den Deepfake [12].

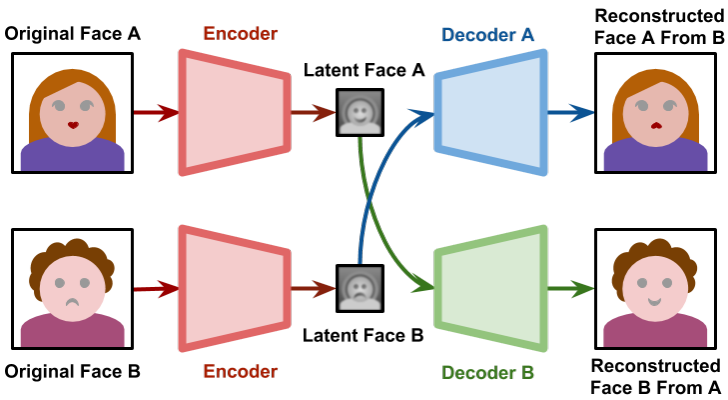

Abbildung 1: Darstellung des Autoencoders zur Erstellung eines Deepfakes [12]

Anwendungsbereiche
Pornoindustrie: Wie bereits erwähnt, wurde 2017 der erste Reddit Blogbeitrag im Bereich Deepfakes und Pornos erstellt. Inzwischen allerdings ist der Subreddit (der Bereich für Deepfakes und Pornografie) aber von Reddit geschlossen worden und das Hochladen selbiger Deepfakes nicht mehr gestattet [13], [14]. 2018 folgte ein weiterer Einsatzbereich von Deepfakes in der Pornoindustrie mit einem neuen Geschäftsmodell. Ein US-Pornostudio veröffentlichte eine Plattform für Kunden, bei welchem man ein eigenes Gesicht hochladen kann und das Studio im Anschluss dieses in einen Pornofilm per Deepfake einbindet. Darüber hinaus kann auch die Umgebung, wie beispielsweise eine Aufnahme der eigenen Wohnung, in das Video animiert werden [15]. Im Juni 2019 ging dann mit DeepNude eine weitere Plattform im Bereich Deepfakes an den Start. Hierbei hat der Nutzer die Möglichkeit, Frauenbilder hochzuladen und im Anschluss ein Nacktbild dieser Fotos herunterzuladen. Inzwischen ist die Anwendung von den Betreibern aber wieder vom Markt genommen worden, da die Angst vor einem Missbrauch des Materials zu groß ist und die App zu

schnell zu viel Aufsehen erregt hat [16], [17]. Es zeigt sich, dass im Bereich Pornografie eine große Nachfrage für die Deepfake Technologie existiert. Dies bestätigt auch eine Studie des niederländischen Unternehmens Deeptrace, die im Oktober 2019 veröffentlicht wurde. Hierbei kam heraus, dass von knapp 15.000 Deepfake Videos, 96% pornografische Videos waren. Vorwiegend seien Deepfakes mit Prominenten entstanden, wie dies auch initial bei Reddit der Fall war [18]. Je besser die Technologie, desto höher auch die Täuschungsgefahr dieser Videos. Dadurch wird auch die Angst vor Missbräuchen, nicht nur bei Prominenten, sondern auch bei Privatpersonen größer. So ereignet sich in Australien der Fall, dass Noelle Martin Opfer eines Deepfakes in diversem Pornomaterial wurde. Als die damals 17-Jährige das Material entdeckt, setzte sie sich im Anschluss erfolgreich für ein Verbot in Australien ein. Das Gesetz wurde 2018 verabschiedet und bestraft nun den Missbrauch von Deepfakes mit einer mehrjährigen Haftstrafe [19], [20].

Filmbranche: Auch in der Filmbranche gibt es diverse Einsatzmöglichkeiten von Deepfakes. Die chinesische Deepfake App „ZAO“ beispielsweise wurde 2019 veröffentlicht und ging innerhalb kürzester Zeit viral. Mithilfe der App war es möglich mittels eines Selfies das eigene Gesicht auf verschiedene Videos per Deepfake zu animieren. Aufgrund fragwürdiger AGBs geriet die App allerdings wieder in Verruf und ist seitdem nicht mehr so beliebt. Nichtsdestotrotz ist es für Nutzer eine einfache und schnelle Variante mit Deepfakes in Berührung zu treten [21]. Ein adäquates Pendant hierzu, ohne die genannten Sicherheitsbedenken, wäre die App „Reface“ [22].
Neben der Unterhaltung haben Deepfakes in der Filmindustrie aber auch einen anderen Nutzen: So kann die Montage eines Gesichts auf einen anderen Schauspieler deutlich vorteilhafter und ggf. auch realistischer durchgeführt werden. Nötig wäre dies zum Beispiel im Film „Fast and Furious 7“ gewesen, als Hauptdarsteller Paul Walker während der Dreharbeiten verstarb. Die Filmcrew hat zur damaligen Zeit allerdings auf die Brüder des verstorbenen zurückgegriffen und mit ihnen die letzten Szenen abgedreht [23].
Auch im Bereich der Lippensynchronisation kann die Filmbranche Deepfakes nutzen, um in verschiedenen Sprachen die Mundbewegungen realistischer darstellen zu können. So muss die Übersetzung des zu sprechenden Textes nicht mehr an die Lippen des Videomaterials angepasst, sondern kann durch die Animation gezeigt werden [24].
Dem Trend der Deepfakes in der Filmbranche verfolgt auch Disney. Der Konzern hat in London auf der Animationsmesse Eurographics Symposium einen Algorithmus zur besseren Übereinanderlegung von Quellmaterialien vorgestellt. Technisch verwendet Disney höher aufgelöste Fotos, bessere Stabilisierungstechniken und automatische Licht- und Kontrastanpassungen für das Training. Dadurch soll der US-Konzern im Vergleich zu anderen Technologien in der Lage sein, Clips mit einer Auflösung von 1024 x 1024 darzustellen. Diese sind deutlich schärfer und werden inzwischen auch als „Megapixel-Deepfakes“ bezeichnet. DeepFaceLab zum Vergleich, kann hingegen nur 256 x 256 Pixel als Output darstellen, da sonst der Effekt des Deepfakes nicht mehr funktioniert. Es zeigt sich zugleich also auch, dass es noch ein langer Weg ist, bis Videomaterialien mit Full HD und Ultra HD per Deepfake dargestellt werden können [25], [26].
Ein weiteres Einsatzgebiet von Deepfakes ist außerdem die Unkenntlichmachung

von Menschen in Videos. Inzwischen gibt es verschiedene Ansätze, die zeigen, dass das klassische Unkenntlich machen von Personen durch Verpixelung des Gesichts keine adäquate Lösung mehr ist, die Identität der Person zu wahren. In Abhängigkeit zum gewählten Verfahren kann so die Erfolgsquote bei bis zu 98% liegen, sodass eine künstliche Intelligenz das Originalbild nach einer Verpixelung wiedererkennen kann [27]. Aufgrund dessen gibt es noch eine andere Möglichkeit die Anonymität zu schützen. Die Schattenwand gilt als ein relativ sicheres Verfahren, kann aber keine Mimiken und Gestiken richtig darstellen, weswegen es aktuell nicht die beste Methode ist. Die HBO Serie „Welcome to Chechnya“ hat daher zum Schutz vor der Verfolgung von Mitgliedern der LBGT-Gemeinde in Tschetschenien diese mittels Deepfakes unkenntlich gemacht. Die Bilderbasis stammt hier von freiwilligen Spenden der eigenen Gesichter, wodurch die Protagonisten nicht mehr identifizierbar, dennoch Emotionen und Gefühle mittels „anderer“ Gesichter darstellbar sind [28].

Kunst: In der Kunstszene können Deepfakes ebenfalls ihren Einsatzzweck finden. Die britische Künstlerin Anna Ridler, welche ihre Kunst bereits im Haus der elektronischen Künste in Basel, aber auch in London oder Linz ausstellt, nutzt Deepfakes für ihre Arbeit. Die Ausstellung „Mosaic Virus“ ist eine Installation von 2019, bei welcher drei Bildschirme verschiedene durch Deepfake generierte Tulpen darstellen. Es soll ein Stillleben des 21. Jahrhunderts darstellen und zeigen, dass Deepfakes auch in der Kunst ihren Mehrwert liefert. Das Trainingsmaterial sammelte die Künstlerin in Amsterdam während der Tulpensaison. Grundlage hierfür waren 20 000 Ablichtungen von Tulpen, die von Hand klassifiziert wurden, um ein neuronales Netz zu trainieren und daraus die künstlichen Tulpen darzustellen [29].

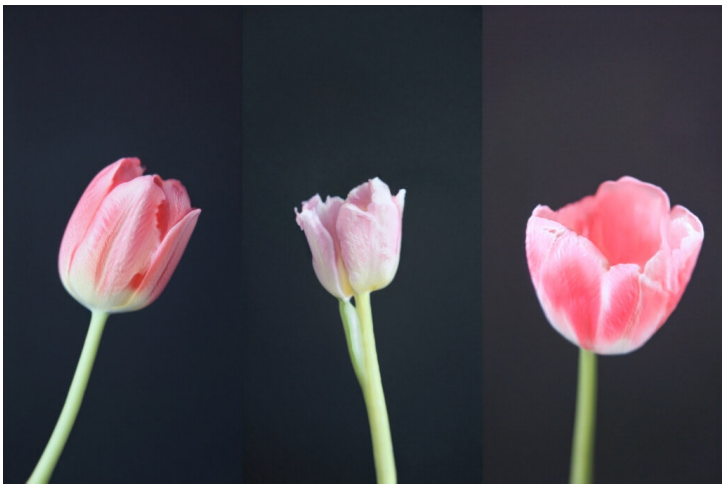

Abbildung 2: Bildinstallation Mosaic Virus [29]

Medizin: Project Revoice ist ein Beispiel für Deepfakes im Bereich der Medizin. Hierbei wird die Technologie verwendet, um chronische Krankheiten, die den Stimmenverlust zur Folge haben, eine authentische Stimme zurückzugeben. Das Projekt bietet verschiedene Arten, vor dem Verlust der Stimme, das Netz zu trainieren. So können beispielsweise Sätze eingesprochen werden, um das neuronale Netz zu trainieren. Im Fall von Pat Quinn, der Co-Founder der Ice Bucket Challenge, war dies nicht mehr möglich. Die Ice Bucket Challenge war eine Spendenaktion, mit dem Zweck, Gelder für ALS erkrankte zu sammeln und zugleich Aufmerksamkeit für die nicht heilbare degenerative Erkrankung zu schaffen. Pat Quinn ist ebenfalls an ALS erkrankt. In seinem Fall war das Einsprechen der Trainingssätze für Project Revoice zu spät. Da er aber in der Öffentlichkeit sehr präsent war und es diverse Videoaufnahmen von ihm gab, haben die Entwickler diese Aufnahmen verwendet, um daraus via Deepfake seine, inzwischen verlorene Stimme, technisch zu reproduzieren. Der Vorteil gegenüber einem Sprachcomputer soll sein, dass das Sprechen sich für den Erkrankten wieder natürlicher anfühlt, da es die eigene Stimme ist, die künstlich über einen Lautsprecher ausgegeben wird [30].

Politik: Der Missbrauch von Deepfakes in sozialen Medien ist relativ groß, weil sich Falschnachrichten dort schnell verbreiten. Vor allem bei einem Medium, wie einem Mobiltelefon, ist es relativ simpel den Zuschauer zu täuschen, ohne eine sonderlich hohe Qualität im Video vorzuweisen. Dafür werden Informationen in sozialen Medien zu schnell konsumiert und die Displays sind zu klein, um zum Beispiel Bildartefakte zu erkennen. Eine Gefahr, die daraus hervorgeht, sind Filterblasen. Sogenannte „Meinungsmache“ kann geschürt werden, indem Nutzer aktiv durch Deepfakes manipuliert werden. Allerdings sagen Experten auch, dass auf der anderen Seite größere Gefahren, wie beispielsweise ein Krieg, nicht durch Deepfakes verursacht werden können. Grund zu der Annahme ist, dass Geheimdienste bereits jetzt mithilfe von aufwändiger Hollywoodtechnik z.B. politische Hetze schüren könnten, dies aber keinen Krieg auslöst [19].
Eine andere Gefahr, die von Deepfakes in dem Kontext Politik ausgeht, ist der Zweifel, an bislang vertrauenswürdigen Quellen, wie Audio- oder Videoaufnahmen. Nachdem eine alte Tonaufnahme von 2005 über Donald Trump, kurz vor den Wahlen, veröffentlicht wurde, in der er sich abfällig über Frauen äußert, hat sich Trump öffentlich entschuldigt. Im Nachgang allerdings hat er gesagt, dass es sich um einen Deepfake handle und daher das Thema für ihn erledigt sei. Ob es nun tatsächlich ein Deepfake war oder nicht bleibt ungeklärt [31]. Ein anderes Beispiel hat sich im Juni 2019 in Malaysia zugetragen. Ein politischer Skandal wurde ausgelöst, als ein Sextape veröffentlicht wurde, welches den Wirtschaftsminister Armin Ali mit einem männlichen rivalisierenden Minister beim Geschlechtsakt zeigt. Da gleichgeschlechtliche sexuelle Handlungen in Malaysia verboten sind, gab es einen öffentlichen Aufschrei. Dieser wurde allerdings im Keim erstickt, da die Regierung behauptet hat, dass es sich um eine Videomanipulation handle, mit dem Ziel die Karriere des Ministers zu sabotieren. Auch wenn unabhängige Wissenschaftler den Deepfake nicht nachweisen konnten und daher dazu plädierten, dass das Video echt sei, konnte so ein öffentlicher Eklat vermieden werden. Wieder stellt sich die Frage, ob bislang belastbare Quellen, wie Audio- und Videoaufnahmen weiterhin verlässlich sind [32].
Identifikation: Auch zur Identifikation und Fälschung von Persönlichkeiten können Deepfakes verwendet werden. Hieraus zeigt sich auch eine der größten Bedenken von Deepfakes von Seiten der Bundesregierung [33]. Video-Ident-Verfahren sind inzwischen ein gängiges Mittel, um sich beispielsweise für ein neues Online-Konto zu authentifizieren. Dabei wird ein Videoanruf, in aller Regel mit einem Dienstleister der Bank, durchgeführt. Der Dienstleister prüft dann, ob das Gesicht mit dem Personalausweis übereinstimmt, ob alle Sicherheitsmerkmale auf dem Personalausweis vorhanden sind und ob die Angaben, mit denen auf dem Ausweis, übereinstimmen. In Folge dessen können Deepfakes verwendet werden, um sich für eine andere Person auszugeben. Beispielsweise könnte ein Angreifer den Personalausweis einer anderen Person stehlen oder fälschen und daraufhin per Video-Ident-Verfahren im Namen der Person mit einem Deepfake exemplarisch ein Konto eröffnen. Da die Videotelefonate in der Regel per Handykamera durchgeführt werden, fällt darüber hinaus auch nicht auf, ob die Qualität des Deepfakes nicht einwandfrei ist [33]. Ein weiteres Verfahren, um sich für eine andere Person auszugeben, ist das Fälschen der Stimme einer Person, zum

Beispiel am Telefon. Dazu hat sich bereits ein Fall im Jahr 2019 ereignet. Cyber Kriminelle haben mithilfe eines Deepfakes den CEO eines Energieunternehmens aus Großbritannien getäuscht. Dieser glaubte mit dem Chef des deutschen Mutterkonzerns zu telefonieren, welcher ihn veranlasste eine Überweisung in Höhe von 220 000€ durchzuführen. Der falsche Chef ordnete an, dass an einen ungarischen Dienstleister der Betrag überwiesen werden soll. Dabei hat er Druck auf den Angestellten ausgeübt, dass das Geld binnen einer Stunde überwiesen sein soll. Nach Angaben des Briten habe sich die Stimme für ihn authentisch angehört, was zeigt, wie weit die Fälschung von Stimmen bereits geht und welche Gefahr beim Identitätsklau mit Deepfakes ausgehen kann [34]. Schließlich können Deepfakes auch zur Tarnung und anonymen Handeln missbraucht werden. So auch der Fall bei Oliver Taylor, ein freier Redakteur mit Schwerpunkt Antisemitismus, der für Zeitungen, wie die Jerusalem Post und die Times of Israel arbeitet. Nachdem er einen Artikel veröffentlicht hat, in dem der Akademiker Mazen Masri und seine Frau Ryvka Barnard, die sich für die Rechte von Palästinensern einsetzt, als “bekannte Sympathisanten des Terrorismus” bezeichnete, fing Reuters an nach dem Autor zu recherchieren. Die Nachrichtenagentur fand heraus, dass dieser Autor nicht existiert. Angaben der Universität, Telefonnummer und E-Mails waren nicht echt und auch sein Profilbild ließ auf einen Deepfake schließen. Da die Zeitungen keine Kontrollmechanismen für Autoren haben, Artikel nur via E-Mail angenommen wurden und Oliver keine Bezahlung forderte, war es für den Autor möglich, eine falsche Identität anzunehmen. Das Deepfake Porträtbild machte das Profil noch glaubwürdiger [35].

Abbildung 3: Portätfoto des Briten Oliver Taylor (links) sowie einer Darstellung der Heatmap des Unternehmens Deepfake Detection Company Cyabra. Gefärbte Bereiche deuten hierbei auf eine Fälschung hin [35].

Allerdings gibt es auch positive Einsatzzwecke im Bereich der Deepfakes und Identitätsschutz: So kann beispielsweise Filmmaterial, welches während einer Testfahrt für das autonome Fahren aufgenommen wurde mit Deepfakes verbessert werden. Zum Schutz der Persönlichkeitsrechte kann die Technik Gesichter aufgenommener Personen austauschen. Ein Trainingsnetz kann infolgedessen trotzdem mit echten menschlichen Gesichtern trainiert werden, trotzdem bleiben Passanten anonym für die künstliche Intelligenz.

Schutzmaßnahmen

Da es einige nützliche Einsatzzwecke von Deepfakes gibt, jedoch auch manche Gefahren von ihnen ausgehen, soll im Folgenden auf Schutzmaßnahmen eingegangen werden. Dabei kann die Gefahr, hervorgerufen durch Deepfakes, auf drei Kernbereiche eingegrenzt werden: Zum einen das Manipulieren von Meinungen durch falsche Inhalte, was vor allem für Social Media eine Gefahr darstellt. Dazu verbunden der Bereich der Politik. Hierbei kann sowohl die Manipulation aber auch die Sabotage ausschlaggebend sein. Letzter Kernbereich ist der Identitätsdiebstahl und die Fälschung einer Identität, da der Mensch bislang davon ausging, dass Quellen, wie

Audio- und Videoaufnahmen als belastbar und verlässlich galten.

Ausgehend vom ersten Kernbereich der Gefahr hat Facebook zusammen mit einigen Partnern, wie Microsoft oder Amazon die „Deepfake Detection Challenge“ kurz: DFDC ins Leben gerufen. Bei der Challenge haben 2114 Teilnehmer insgesamt 35 000 KI-Modelle eingereicht. Ziel war Deepfakes von realen Videos zu unterscheiden, wobei die Modelle zuvor nicht mit dem Material trainiert wurden. Es handelte sich folglich um „ungesehenes“ Material, welches mehr als 100 000 Videos umfasste. Um die Erkennung noch schwieriger zu machen, enthielten die Videos darüber hinaus Elemente, die die Erkennungssysteme verwirren sollten: Beispielsweise ein Makeup-Tutorial, bei der eine Person geschminkt wird, Videos, welche durch Einfügen von Text und Formen über den Gesichtern der Sprecher ablenken sollten, das Ändern von Auflösung oder der Ausrichtung sowie langsameres und schnelleres Abspielen eines Videos. Das Modell erlangte hierbei eine Accuracy von 65%. In Folge der Nominierung wurde gesagt, dass das beste Modell sehr gut, in Anbetracht der schwierigen Bedingungen agiert hat, jedoch das System trotzdem nicht gut genug sei, um auf den Social Media Plattformen dauerhaft zum Einsatz zu kommen [36], [37].

Während der US-Wahlen 2020 sollen Deepfakes von den Plattformen jedoch mithilfe dieser Techniken gesondert beobachtet und verbannt werden. Hierfür haben Facebook, YouTube und Twitter bereits Anfang des Jahres ihr Regelwerk für Anti-Deepfake-Paragraphen erweitert. Demnach sind satirische und parodistische Inhalte weiterhin erlaubt, manipulative Inhalte jedoch verboten [38].
Auch in der Politiksparte gibt es Schutzmaßnahmen. Neben beispielsweise dem Gesetz zum Verbot von Deepfakes in Australien wurde im Oktober 2019 in Kalifornien für 60 Tage das Verbreiten von „materially deceptive audio or visual media“ in Bezug auf die Kandidatur, kurz vor der Wahl, verboten [33]. In Deutschland hingegen ist dieser Rechtsraum allerdings nicht so eindeutig gesetzlich geregelt.

Ein weiterer denkbarer Lösungsweg wäre die verpflichtende Einbindung eines digitalen Wasserzeichens, welches Rückschlüsse über die Produktion des Videos und die Verbreitungswege, beispielsweise durch die Blockchain Technologie, ermöglichen könnte [39].

Im Bereich des Identitätsklaus gibt es inzwischen auch Schutzmaßnahmen. Vor allem bei Audioaufnahmen gibt es verschiedene Verfahren, um Fälschungen durch Deepfakes von echten zu unterscheiden: So hat das Frauenhofer-Institut eine künstliche Intelligenz entwickelt, welche es ermöglicht den Unterschied zwischen Textbausteinen und gesprochener Stimme zu erkennen [40]. Außerdem können Schwankungen in den Netzfrequenzen anhand von Audioaufnahmen identifiziert werden. Da das Stromnetz nicht immer exakt 50 Hertz liefert, kann in den Audioaufnahmen das Aufzeichnungsdatum abgeglichen werden. Da diese Audiocharakteristiken über Jahre aufgezeichnet wurden, kann so die Validitätsprüfung stattfinden [39].

Erkennungssoftware und deren Manipulation

Eine Erkennungssoftware, die im Gegensatz zu den Modellen der DFDC mit bereits gesehenem Material trainiert wurde, nennt sich FaceForensics++ und wurde im Jahr 2019 vorgestellt. Als Datenbasis wurden hierbei YouTube Videos verwendet, welche mit verschiedenen Techniken, wie FaceSwap

oder Face2Face manipuliert und im Anschluss mit drei verschiedenen Videoqualitäten von der Software untersucht wurden. Eine dazu passende Nutzerstudie hat gezeigt, dass vor allem mit sinkender Qualität die Erkennung durch das menschliche Auge deutlich schlechter wird. FaceForensics++ hingegen kann hier noch deutlich bessere Ergebnisse erzielen. Dazu wurde ein automatischer Classifier entwickelt, welcher auf Grundlage von Artefakten in den Bildern, neben dem gewählten Manipulationsmodell (Face2Face, FaceSwap, usw.) auch klassifizieren konnte, ob es sich um eine Fälschung handelt oder nicht. Darüber hinaus kann das Modell auch feststellen, an welcher Stelle im Bild die Manipulation zum Einsatz kommt [41].

Ein Jahr später folgte das Paper „Adversarial Deepfakes“, welches sich zum Ziel gesetzt hat, die Erkennung von z.B. FaceForensics++ wiederrum zu täuschen. Dazu wird als Grundlage eine Adversarial Attack verwendet, die auf einem normalen Bild ein künstlich erzeugtes Rauschen hinzuaddiert. Für das menschliche Auge sieht das Bild immer noch gleich aus, jedoch extrahiert die künstliche Intelligenz aufgrund des Rauschens andere Kernmerkmale aus dem Bild, was anschließend zu einer Fehlinterpretation führt [42].

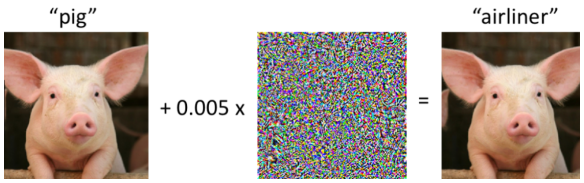

Abbildung 4: Beispiel einer Adversarial Attack - Links ein Bild eines Schweins, welches nach dem Hinzufügen des Rauschens von der KI als "airliner" klassifiziert wird. Für das menschliche Auge gibt es keinen Unterschied zu erkennen [43].

Durch diese Angriffsmethode können folglich automatische Erkennungen getäuscht werden.

Die Erfolgsrate unterscheidet sich dahingehend, ob es sich um eine Whitebox-Attacke oder eine Blackbox-Attacke handelt und ob das Video vor der Erkennung komprimiert wird. Im Fall der Whitebox-Attacke ist die Erkennung meist am einfachsten, da die Architektur des Detektors sowie die Arbeitsweise und verwendete Parameter bekannt sind. In diesem Fall liegt die Erkennung bei nahezu 100%. Bei einer Blackbox-Attacke sinkt diese in Abhängigkeit zur gewählten Erkennung um 5 bzw. 15 Prozentpunkte. Wird allerdings das Video komprimiert, liegt die Täuschungsrate in Abhängigkeit zum Erkennungsmodell nur noch bei 58% bzw. 92% [42].

Es zeigt sich also, dass das Wechselspiel zwischen Erkennung und Täuschung mithilfe von KI im Bereich Deepfakes über die Zeit immer weiter vorangetrieben wird.

Fazit

Festzuhalten ist, dass es nicht nur negative Seiten von Deepfakes gibt, wie es häufig von den Medien dargestellt wird. Vor allem die positiven Seiten von Deepfakes können einen Mehrwert für die verschiedensten Bereiche bringen. Dabei zeigt sich, dass mit fortschreitender Technik die Entwicklung von Deepfakes zunimmt. Der Aufwand wird immer geringer Deepfakes zu erzeugen und zugleich wird die Technik immer leichter zugänglich für Privatpersonen. Außerdem ist auch ein Trend zu verzeichnen, dass die Qualität immer besser wird. Vor allem im Bereich der Filmbranche kann eine Verbesserung der Qualität auf Full HD bzw. Ultra HD einen großen Innovationsschritt bedeuten, da aufwändige Effekte deutlich schneller und kostensparender umgesetzt werden können.
Trotzdem sollte man auch die Gefahren im Bereich der Social Media nicht außer Acht

lassen. Es ist wichtig, dass weiterhin an Erkennungsmechanismen geforscht wird, um die Nutzer zu schützen. Darüber hinaus sollte auch die Gesetzeslage in Deutschland überarbeitet werden. Bislang fehlt es hier an einer einheitlichen gesetzlichen Regelung, die eindeutig eine Grenze zwischen einer zulässigen Bearbeitung von Videos und deren Verbreitung und unzulässigen Täuschungen ziehen kann. Dies liefert zwar keinen vollumfänglichen Schutz vor Deepfakes und Fake News, kann aber eine Absicherung sein. Schließlich sollte auch in der Gesellschaft das Misstrauen gegenüber Video- und Audiodateien größer werden. Es muss eine Aufklärung stattfinden, um zu zeigen, wie weit die Technik heutzutage ist, um bislang verlässlichen Audio- und Videoquellen mehr Skepsis entgegenzubringen.

Quellen

[1] „Infografik: Künstliche Intelligenz, Chance oder Riskio?“, *Statista Infografiken*. https://de.statista.com/infografik/16394/einsatzgebiet-von-kuenstlicher-intelligenz/ (zugegriffen Sep. 16, 2020).

[2] „Was ist ein Deepfake?“ https://www.bigdata-insider.de/was-ist-ein-deepfake-a-915237/ (zugegriffen Sep. 22, 2020).

[3] C. Bregler, M. Covell, und M. Slaney, „Video Rewrite: driving visual speech with audio“, in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*, Not Known, 1997, S. 353–360, doi: 10.1145/258734.258880.

[4] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, und M. Nießner, „Face2Face: Real-time Face Capture and Reenactment of RGB Videos“.

[5] S. Suwajanakorn, S. M. Seitz, und I. Kemelmacher-Shlizerman, „Synthesizing Obama: learning lip sync from audio“, *ACM Trans. Graph.*, Bd. 36, Nr. 4, S. 1–13, Juli 2017, doi: 10.1145/3072959.3073640.

[6] J. Vincent, „New AI research makes it easier to create fake footage of someone speaking“, *The Verge*, Juli 12, 2017. https://www.theverge.com/2017/7/12/15957844/ai-fake-video-audio-speech-obama (zugegriffen Sep. 22, 2020).

[7] „Deepfakes“, *IONOS Digitalguide*. https://www.ionos.de/digitalguide/online-marketing/social-media/deepfakes/ (zugegriffen Sep. 22, 2020).

[8] S. N. Germany Stuttgart, „Video-Manipulation im Internet: Deepfakes – Fake News auf dem nächsten Level?“, *stuttgarter-nachrichten.de*. https://www.stuttgarter-nachrichten.de/inhalt.video-manipulation-im-internet-deepfakes-fake-news-auf-dem-naechsten-level.9a5c74bd-450a-40b8-8cc1-f9d47306132b.html (zugegriffen Sep. 22, 2020).

[9] „Deepfakes FakeApp“. https://www.heise.de/download/product/deepfakes-fakeapp (zugegriffen Sep. 22, 2020).

[10] iperov, „https://github.com/iperov/DeepFaceLab“, Sep. 22, 2020. https://github.com/iperov/DeepFaceLab (zugegriffen Sep. 22, 2020).

[11] E. Zakharov, A. Shysheya, E. Burkov, und V. Lempitsky, „Few-Shot Adversarial Learning of Realistic Neural Talking Head Models“, *ArXiv190508233 Cs*, Mai 2019, Zugegriffen: Sep. 22, 2020. [Online]. Verfügbar unter: http://arxiv.org/abs/1905.08233.

[12] A. Zucconi, „Understanding the Technology Behind DeepFakes“, *Alan Zucconi*, März 14, 2018. https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/ (zugegriffen Sep. 22, 2020).

[13] „r/announcements - Update on site-wide rules regarding involuntary pornography and the sexualization of minors“, *reddit*.

https://www.reddit.com/r/announcements/comments/7vxzrb/update_on_sitewide_rules_regarding_involuntary/ (zugegriffen Sep. 22, 2020).
[14] „Reddit bans 'deepfakes' AI porn communities - The Verge". https://www.theverge.com/2018/2/7/16982046/reddit-deepfakes-ai-celebrity-face-swap-porn-community-ban (zugegriffen Sep. 22, 2020).
[15] J. Roettgers und J. Roettgers, „Naughty America Wants to Monetize Deepfake Porn", *Variety*, Aug. 20, 2018. https://variety.com/2018/digital/news/deepfake-porn-custom-clips-naughty-america-1202910584/ (zugegriffen Sep. 22, 2020).
[16] „Creator of DeepNude, App That Undresses Photos of Women, Takes It Offline". https://www.vice.com/en_us/article/qv7agw/deepnude-app-that-undresses-photos-of-women-takes-it-offline (zugegriffen Sep. 22, 2020).
[17] „deepnudeapp (@deepnudeapp) / Twitter", *Twitter*. https://twitter.com/deepnudeapp (zugegriffen Sep. 22, 2020).
[18] „Mapping the Deepfake Landscape", *Sensity*, Okt. 07, 2019. https://sensity.ai/mapping-the-deepfake-landscape/ (zugegriffen Sep. 22, 2020).
[19] „Künstliche Intelligenz: Deepfakes – darum sollten wir besorgt sein". https://www.zdf.de/uri/d2399b9f-ba10-480c-be90-7108e65d0db1 (zugegriffen Sep. 22, 2020).
[20] A. told to D. Scott, „Deepfake Porn Nearly Ruined My Life", *ELLE*, Feb. 06, 2020. https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/ (zugegriffen Sep. 22, 2020).
[21] „Chinesische Deepfake-App sorgt für Aufregung – die User-Videos sind krass!", *watson.ch*. https://www.watson.ch/!549945806 (zugegriffen Sep. 22, 2020).
[22] „REFACE: Gesichter austauschen in Foto und Videos – Apps bei Google Play". https://play.google.com/store/apps/details?id=video.reface.app&hl=de (zugegriffen Sep. 22, 2020).
[23] S. Zeitung, „‚Fast and Furious' Paul Walkers Brüder bringen Film zu Ende", *Süddeutsche.de*. https://www.sueddeutsche.de/kultur/fast-and-furious-7-brueder-springen-fuer-paul-walker-ein-1.1938770 (zugegriffen Sep. 22, 2020).
[24] „Deepfakes: Sind manipulierte Videos eine Gefahr?", *MAZ - Märkische Allgemeine*. https://www.maz-online.de/Nachrichten/Digital/Deepfakes-Sind-manipulierte-Videos-eine-Gefahr (zugegriffen Sep. 22, 2020).
[25] „Deepfakes bei Disney und in Dokus: Die Porno-Technologie ist in Hollywood angekommen", *stern.de*. https://www.stern.de/digital/online/deepfakes-bei-disney-und-in-dokus--die-porno-technologie-ist-in-hollywood-angekommen-9320088.html (zugegriffen Sep. 22, 2020).
[26] M. Schreiner, „Disney entwickelt Megapixel-Deepfakes fürs Kino", *MIXED | News zu VR, AR und KI*, Juli 01, 2020. https://mixed.de/disney-entwickelt-megapixel-deepfakes-fuers-kino/ (zugegriffen Sep. 24, 2020).
[27] V. Tischbein, „Verpixelung macht unsichtbar - oder doch nicht?", *netzpolitik.org*, Sep. 15, 2016. https://netzpolitik.org/2016/verpixelung-macht-unsichtbar-oder-doch-nicht/ (zugegriffen Sep. 22, 2020).
[28] R. Heilweil, „How deepfakes could actually do some good", *Vox*, Juni 29, 2020. https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya (zugegriffen Sep. 22, 2020).
[29] „Mosaic Virus, 2019", *ANNA RIDLER*. http://annaridler.com/mosaic-virus (zugegriffen Sep. 22, 2020).
[30] „Home - Project Revoice".

https://www.projectrevoice.org/ (zugegriffen Sep. 22, 2020).
[31] M. Haberman und J. Martin, „Trump Once Said the 'Access Hollywood' Tape Was Real. Now He's Not Sure.", *The New York Times*, Nov. 28, 2017.
[32] N. Ker, „Is the political aide viral sex video confession real or a Deepfake? | Malay Mail". https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-a-deepfake/1761422 (zugegriffen Sep. 22, 2020).
[33] „Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Manuel Höferlin, Frank Sitta, Grigorios Aggelidis, weiterer Abgeordneter und der Fraktion der FDP", 19/15657, Dez. 2019.
[34] C. Stupp, „Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case", *Wall Street Journal*, Aug. 30, 2019.
[35] R. Satter, „Deepfake used to attack activist couple shows new disinformation frontier", *Reuters*, Juli 15, 2020.
[36] „Deepfake Detection Challenge Dataset". https://ai.facebook.com/datasets/dfdc (zugegriffen Sep. 24, 2020).
[37] B. Dolhansky *u. a.*, „The DeepFake Detection Challenge Dataset", *ArXiv200607397 Cs*, Juni 2020, Zugegriffen: Sep. 24, 2020. [Online]. Verfügbar unter: http://arxiv.org/abs/2006.07397.
[38] M. Bastian, „TikTok: Deepfake-Bann zur US-Wahl 2020", *MIXED | News zu VR, AR und KI*, Aug. 05, 2020. https://mixed.de/tiktok-deepfake-bann-zur-us-wahl-2020/ (zugegriffen Sep. 24, 2020).
[39] N. Lossau, „Gefahren, Herausforderungen", S. 9, 2020.
[40] „Kampfansage an Deepfakes - Fraunhofer IDMT", *Fraunhofer-Institut für Digitale Medientechnologie IDMT*. https://www.idmt.fraunhofer.de/de/Press_and_Media/press_releases/2019/kampf-gegen-deepfakes.html (zugegriffen Sep. 24, 2020).
[41] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, und M. Nießner, „FaceForensics++: Learning to Detect Manipulated Facial Images", *ArXiv190108971 Cs*, Aug. 2019, Zugegriffen: Sep. 24, 2020. [Online]. Verfügbar unter: http://arxiv.org/abs/1901.08971.
[42] P. Neekhara, S. Hussain, M. Jere, F. Koushanfar, und J. McAuley, „Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples", *ArXiv200212749 Cs*, März 2020, Zugegriffen: Sep. 24, 2020. [Online]. Verfügbar unter: http://arxiv.org/abs/2002.12749.
[43] K. S. Sivamani, „The unusual effectiveness of adversarial attacks", *Medium*, Juli 31, 2019. https://medium.com/@smkirthishankar/the-unusual-effectiveness-of-adversarial-attacks-e1314d0fa4d3 (zugegriffen Sep. 24, 2020).