Deepfakes – Perfekt gefälschte Bilder und Videos

Zusammenfassung

Gefälschte Bilder sind nichts Neues. Aber die Fähigkeit, alternative Fakten zu schaffen hat mit der Deepfake-Technologie einen signifikanten Sprung gemacht. Es ist mittlerweile relativ einfach möglich, Audio- und Video-Dateien von echt wirkenden Menschen zu erstellen, die Dinge sagen und tun, die sie nie gesagt oder getan haben. Dabei werden eigenständig lernende Algorithmen wie neurale Netzwerke mit Audio und Bildbearbeitungssoftware kombiniert. Das Ergebnis sind – für den/die Laien/in nicht vom Original zu unterscheidende – Fälschungen. Welche Risiken birgt das? Vor allem Videos, die als Medium bisher ein relativ hohes Vertrauen genießen, können für Rufschädigung, Erpressung oder Marktmanipulation missbraucht werden – mit eklatanten Folgen für Einzelne, Unternehmen und Gesellschaft. Potenzielle Gefahr für die Demokratie, sowie nationale und internationale Sicherheit könnte etwa von gefälschten Aufnahmen von Gewalt, Kriegserklärungen, Ankündigungen drohender Katastrophen oder Beweisen für das kriminelle Verhalten eines Staatsoberhaupts ausgehen. Auch Wahlbeeinflussung oder staatlicher, z. B. polizeilicher oder geheimdienstlicher Missbrauch der Technologie sind denkbar. Außerdem könnte die unkontrollierte Verbreitung von Deepfakes zu einem starken Verlust von Vertrauen in politische und mediale Institutionen führen.

Überblick zum Thema

Nachdem verändertes und gefälschtes Foto- und Videomaterial so alt ist wie die Aufnahmetechniken selbst, hat der Realismus und die Einfachheit des mit Künstlicher Intelligenz „gefakten" Bildes und Tons eine neue Stufe erreicht. Es ist mittlerweile mit frei erhältlichen Apps und Programmen möglich, komplexes Material wie z. B. Gesicht und Sprache einer Person nachzuahmen und in für Laien/innen echt erscheinendes Videomaterial zu verwandeln. Solche „Deepfakes" haben jüngst viel Aufmerksamkeit auf sich gezogen, z.B. durch komplett gefälschte Reden von PolitikerInnen[1] oder gefälschte Pornographie mit berühmten Persönlichkeiten.[2] Der Begriff Deepfake setzt sich dabei aus „deep learning", einer Methode, den Lernerfolg künstlicher neuronaler Netze zu optimieren, und „fake", also Fälschung, zusammen.

Künstliche Intelligenz kombiniert mit Audio- und Bildbearbeitungssoftware

Bei früheren Fälschungen wurde beispielsweise das Gesicht einer Person aus vorhandenem Bildmaterial ausgeschnitten und über das Gesicht einer anderen montiert. Der Prozess war relativ aufwendig, die Ergebnisse insbesondere bei bewegten Bildern oft offensichtlich zweifelhaft. Fälschungen waren relativ leicht an Übergangsstellen oder an Beleuchtungswinkeln zu

[1] You Won't Believe What Obama Says In This Video! youtu.be/cQ54GDm1eL0.

[2] Künstliche Intelligenz: Auf Fake News folgt Fake Porn, *Die Zeit,* zeit.de/digital/internet/2018-01/kuenstliche-intelligenz-deepfakes-porno-face-swap.

erkennen. Solche „face swaps“ sind mittlerweile in spielerischen Apps weitverbreitet und liefern nahezu ohne Verzögerung auf den ersten Blick gute Ergebnisse;[3] ähnliche Apps generieren Fake-Video-Anrufe[4] oder verändern den Körper in gewünschter Weise.[5] Fälschungen, die mit Hilfe Künstlicher Intelligenz, größerer Rechenleistung und professioneller Software erzeugt werden, sind dagegen noch viel überzeugender, weil hier 3D-Computergrafikmodelle, z. B. des gefälschten Gesichts, von Grund auf generiert werden. Hier entstehen bereits die ersten Geschäftsmodelle, welche den Zugang zu professioneller Deepfake-Software und -Rechenleistung stundenweise verkaufen.[6]

Maschinelles Lernen: Neuronale Netzwerke trainieren sich gegenseitig

Oft kommen dabei zwei gegnerische neuronale Netzwerke zum Einsatz, die selbstständig voneinander lernen (generative adversary networks, GAN, vgl. Goodfellow/Pouget-Abadie et al. 2014). Einer dieser lernenden Algorithmen (Generator) wird mit mehreren Stunden vorhandenen Videomaterials mit dem Ziel trainiert, anhand von vielen Variablen möglichst genaue Kopien erzeugen zu können. Der zweite Algorithmus, der Diskriminator, wird dahingegen trainiert, die Ergebnisse des ersten vom Original zu unterscheiden. Der Generator versucht Ergebnisse, also z. B. das Modell eines sprechenden Menschen, zu erzeugen, die der Diskriminator nicht mehr unterscheiden kann. Dadurch nähern sich die gefakten Inhalte dem Erscheinungsbild des Originals nach und nach immer weiter an.

perfekt gefälschte Mimik und Stimme

Hat der erste Algorithmus alle nötigen biometrischen Parameter und Eigenheiten wie Mimik, Mundbewegungen und Sprache einer Person in verschiedenen Situationen erlernt, kann professionelle Audio- und Bildverarbeitungssoftware z. B. ein Gesicht perfekt digital replizieren und in ein beliebiges Video derselben oder einer anderen Person in hoher Qualität einfügen oder ein neues Video erzeugen. Da auch alle charakteristischen Stimmeigenschaften wie Frequenz, Intonation oder Pausen erlernt und digital repliziert werden, passen dann nicht nur die Lippenbewegungen perfekt zum vermeintlich Gesagten, sondern auch die Stimme selbst. Dafür reichen schon wenigen Minuten gesprochenen Materials.

Erkennung von Fälschungen

Für den/die Laien/in nicht zu erkennen, haben ExpertInnen verschiedene Ansätze entwickelt, um gefälschtes Bildmaterial zu entlarven. Beispielsweise fehlt in künstlich erzeugten Videos oft das physiologisch wichtige Augenblinzeln, dies machen sich ForscherInnen zunutze, um wiederum

---

[3] Dazu gibt es mittlerweile eine Reihe von frei verfügbaren Apss, siehe play.google.com/store/search?q=deepfake+app&c=apps&hl=de&gl=US.

[4] Im Juni 2022 wurde über Videotelefonate zwischen einem vermutlich deep-gefakten Kiewer Bürgermeister Vitali Klitschko mit europäischen BürgermeisterInnen, darunter der Wiener Michael Ludwig berichtet; unklar ist, ob in diesem Fall technisch oder mit einem Schauspieler gefakt wurde. Offenbar ist aber die Technik schon so weit fortgeschritten, dass nur mehr an der Plausibilität der gesprochenen Inhalte erkennbar ist, mit wem man es zu tun hat. Siehe futurezone.at/digital-life/wiener-buergermeister-michael-ludwig-deepfake-vitali-klitschko-ukraine-krieg-giffey/402053326.

[5] Gesichter spielerisch tauschen mit Apps: play.google.com/store/search?q=deepfake&c=apps&hl=de.

[6] deepfakesweb.com.

lernende Algorithmen auf die Erkennung solcher Abnormalitäten zu trainieren (Li, Chang et al. 2018). Auch unsichtbare „Wasserzeichen“ die auf eine bestimmte Kamera zurückzuführen sind, sind in Planung, oder auch Blockchain basierte verifizierbare Zeitmarken. Lernende Algorithmen wurden auch benutzt, um Kunstfälschungen zu entlarven, da alle Charakteristika jedes Pinselstriches des Gesamtwerks eines/r Künstlers/in analysiert und Abweichungen erkannt werden können. In einem Fall produzierte ein 3D-Drucker auf Basis solcher Daten ein Portrait Rembrandts das alle Charakteristika eines echten Rembrandts aufweist, eine künstliche Intelligenz würde dieses vermutlich als echt klassifizieren (Floridi 2018). Auch das laufende österreichische Projekt Defalsif-AI[7] des AIT beschäftigt sich mit KI-basierter Software zur Detektion von Deepfakes.

Wettlauf zwischen EntwicklerInnen

Offen ist dabei die Frage, wie schnell die Entwicklung von Fälschungstechnologie auf Erkennungen reagiert. Während 2018 fehlendes Augenblinzeln als unfehlbares Kennzeichen für Deepfakes galt, tauchten wenig später die ersten Videos mit Augenblinzeln auf. Der Wettlauf zwischen EntwicklerInnen ist in vollem Gange und es ist nicht immer klar, wer gerade die Oberhand hat, und auf welcher Seite er/sie steht. Derzeit gelten Blockchain-Technologien als hoffnungsvollster Weg die Authentizität von Videos zweifelsfrei zu dokumentieren (Hasan/Salah 2019), und KI-Systeme selbst werden zur Detektion von Fakes eingesetzt. Dass KI auch zu solchen Zwecken genutzt wird, ist offensichtlich und verschiedene Formen der Kriminalität treten heute entweder schon auf oder sind in naher Zeit absehbar (King, Aggarwal et al. 2018, siehe auch Thema „Künstliche Intelligenz“).

Schäden für Personen und Unternehmen; Marktmanipulation

In Bezug auf Deepfakes stehen momentan strafbare Handlungen gegen Personen im Vordergrund. Die oben beschriebenen gefälschten Bilder und Videos können nicht nur rufschädigend wirken und Belästigungen hervorrufen, sondern bei Betroffenen auch zu schweren psychologischen Auswirkungen führen, insbesondere falls einzelne Fälle von den Medien ausführlich aufgegriffen werden. Auch Erpressungen und Identitätsdiebstähle könnten sich häufen, was z. B. auch eklatante Auswirkungen auf Unternehmen haben könnte. Auch Marktmanipulationen, beispielsweise am Aktienmarkt, sind mit gefakten Aussagen von CEOs denkbar. Auch Videoanrufe mit dem Gesicht des CEOs, der dann die gefälschte Anweisung erteilt, Geld auf ein bestimmtes Konto zu überweisen, sind bereits bekannt geworden (CEO-Fraud). Besonders in den letzten Jahren, in denen wegen erhöhter Ansteckungsgefahr viele Besprechungen auf Videoplattformen verschoben wurden, erlebten Deep Fakes eine Blüte.

Rufschädigung, Belästigung, Einschüchterung, Erpressung

Momentan sind vor allem Persönlichkeiten öffentlichen Lebens betroffen, da für einen guten Fake genug Trainingsmaterial in Form von Bildern und Videos im Internet verfügbar sein muss. Allerdings reichen für einen täuschend echten „faceswap“, also den Austausch eines Gesichts, schon wenige Hundert Bilder der betreffenden Person. Durch den allgegenwärtigen Gebrauch von Smartphones und die weitverbreitete Speicherung von Fotos

[7] ait.ac.at/themen/surveillance-protection/projects/defalsif-ai.

in angreifbaren Clouds oder in Sozialen Netzwerken kann potenziell jedeR NutzerIn zum Opfer werden (siehe Thema „Cloud Computing“). Besonders Selfies eigenen sich gut als Trainingsmaterial für die Neuronalen Netzwerke. Die einfache Fälschbarkeit von bewegtem Bild und Stimme könnten auch weitreichende Auswirkungen auf Nutzung und Missbrauch von sprachgesteuerten Geräten und Systemen mit biometrischer Fernauthentifizierung haben. Weiters werden Deepfakes in Videoanrufen immer mehr auch zu einem Sicherheitsproblem in Organisationen. Gezieltes Social Engineering, sog. Spear-Phishing, bedient sich mittlerweile auch der Deepfakes.[8] Wenn sich Missbrauchsfälle insgesamt häufen, kann das zu einem gravierenden gesellschaftlichen Problem werden.

Relevanz des Themas für das Parlament und für Österreich

potenzielle Gefahr für Demokratie und nationale Sicherheit

Demokratiegefährdend wird die Entwicklung dann, wenn nicht Prominente, wie z. B. Filmstars, Opfer von Deepfakes werden, sondern EntscheidungsträgerInnen und PolitikerInnen. Das könnte auch Risiken für die nationale und internationale Sicherheit bergen (Chesney and Citron 2018). Gefälschtes Bildmaterial kann als digitaler Beweis für beliebige Situationen genutzt werden. Nach dem gleichen Prinzip wie „Fake News“ spielt die Authentizität eines Videos oft keine Rolle mehr, nachdem es mehrfach über Social Media geteilt wurde (siehe Thema „Robojournalismus“, und „Microtargeting“). Aufwendige Dementi-Kampagnen binden dann nicht nur Ressourcen, sondern führen auch oft durch die erhöhte mediale Aufmerksamkeit zu einer noch weiteren Verbreitung der Fälschungen. Bislang war Österreich noch nicht stark von diesem Phänomen betroffen, doch in absehbarer Zeit könnte sich das ändern, es gibt jedenfalls keine Barriere die eine Ausbreitung in Österreich verhindern könnte.

Auswirkungen auf Wahlen, Schüren von Unruhen

Tatsächlich wird eine gefälschte Rede eines/r Kandidaten/in bei einer politischen Wahl, ob als solche identifiziert oder nicht, wahrscheinlich Auswirkungen auf seine/ihre WählerInnenschaft haben. Gefälschte Aufnahmen von Polizeigewalt, Kriegserklärungen, Ankündigungen drohender Katastrophen oder Beweise für das kriminelle Verhalten eines Staatsoberhaupts haben das Potenzial, unmittelbar zu sozialen Unruhen zu führen (EPRS 2018). Auch staatlicher, z. B. polizeilicher oder geheimdienstlicher, Missbrauch der Technologie ist denkbar, etwa zur Beweismittelfälschung in illiberalen Regimen.

möglicher Verlust von Vertrauen in Institutionen

Ob professionell gefälschte Videos zum Massenphänomen werden, ist derzeit vielleicht noch fraglich, da die Technologie dafür aber reif und einfach verfügbar ist, steht dem wenig entgegen. Gesellschaftliches Bewusstsein für die nahezu perfekte Fälschbarkeit von Video- und Audiomaterialien zu schaffen, ist ein demokratiestärkender Ansatz. Aber auch das weit verbreitete Bewusstsein darüber könnte selbst zu einem gravierenden Problem werden, weil das ohnehin schon angeschlagene Vertrauen in öffentliche In-

[8] computerweekly.com/de/tipp/Phishing-mit-Deepfakes-Unternehmen-muessen-sich-vorbereiten.

stitutionen[9] und Medien dadurch noch weiter ausgehöhlt werden könnte. Wenn jedes Video ein Fake sein könnte, was und wem glaubt man dann?

Vorschlag für weiteres Vorgehen

Eine genaue Erhebung des technischen Ist-Zustandes und dessen Weiterentwicklungspotenzials wäre die Grundlage für eine tiefergreifende Abschätzung und Bewertung bisheriger sowie möglicher sozialer, politischer und wirtschaftlicher Folgen. Auf dieser wissenschaftlichen Basis könnten dann Empfehlungen für weitere Maßnahmen zum gesetzlichen, institutionellen und organisatorischen Umgang mit Deepfakes erarbeitet werden. Vorzugsweise könnte das unter Einbindung von Stakeholdern, ExpertInnen und einer breiten Öffentlichkeit geschehen. Das Thema sollte dringend behandelt werden, um mögliche Schäden von Demokratie und öffentlichem Leben abzuwenden. Im Mai 2022 hat die Bundesregierung in Erfüllung einer Entschließung des Parlaments[10] einen Nationalen Aktionsplan Deepfake (BKA et al. 2022) beschlossen.[11] Dieser Plan zitiert u.a. die Erstfassung des vorliegenden Themenpapiers, erörtert die Bedrohungslage und beschreibt Maßnahmen in mehreren Handlungsfeldern.

Nationaler Aktionsplan Deepfake 2022

Zentrale weiterführende Quellen

BKA, BMEIA, BMJ, BMLV und BMI, 2022, Aktionsplan Deepfake, Wien, bmi.gv.at/bmi_documents/2779.pdf.

BSI (Bundesamt für Sicherheit in der Informationstechnik), o.J., Deepfakes – Gefahren und Gegenmaßnahmen, bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html.

Chesney, B., D. Citron, 2018, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. Univ. of Texas School of Law.

EPRS, 2018, Global Trendometer. Essays on medium- and long-term global trends. Brussels, European Parliamentary Research Service - European Parliament.

Floridi, L., 2018, Artificial Intelligence, Deepfakes and a Future of Ectypes. Philosophy & Technology 31(3): 317-321.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, 2014, Generative adversarial nets. Advances in neural information processing systems: 2672–2680.

King, T., N. Aggarwal, M. Taddeo and L. Floridi, 2018, Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions, Oxford Internet Institute, University of Oxford.

Li, Y., et al., 2018, In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. (preprint) arXiv:1806.02877."

Hasan, H. R. und Salah, K., 2019, Combating Deepfake Videos Using Blockchain and Smart Contracts, *IEEE Access* 7, 41596-41606.

[9] Standard Eurobarometer 88 – Public opinion in the European Union, ec.europa.eu/commfrontoffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/82873.

[10] Entschließung Nr. 104/E vom 14. Oktober 2020, parlament.gv.at/PAKT/VHG/XXVII/E/E_00104/.

[11] bmi.gv.at/bmi_documents/2779.pdf.