W08_Ethik_Deepfakes

Deepfakes – der Missbrauch von KI

Erarbeitet von
Dr. Jacqueline Klusik-Eckert

Lernziele

- Erhalten eine Sensibilisierung für den Missbrauch von KI
- Lernen die Definition von Deepfake kennen
- Können zwischen dem Oberbegriff und dem Face-Reenactment differenzieren
- Verstehen das moralische und ethische Dilemma von Deepfake
- Wissen um die Möglichkeiten, Deepfakes zu erkennen

Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Wie gut können wir unseren Augen eigentlich noch trauen? Oder besser: Wie authentisch sind Videos im Netz?
Da schneidet auf einmal die Mona Lisa lustige Grimassen.

Quelle [1]

Tom Cruise blödelt auf TikTok herum.

Quelle [2]

John F. Kennedy macht Kinderreime.

Quelle [3]

oder Barack Obama beschimpft Donald Trump.

Quelle [4]

Das ist doch alles nur Spaß. Wirklich?
Solange die Mona Lisa lustige Grimassen schneidet, ist alles ok. Jeder weiß sofort, dass es sich um eine Manipulation am Video gehandelt hat. Aber wie sieht es in den anderen Fällen aus?
Was sind die ethischen Konsequenzen, wenn wir jede Person alles sagen und machen lassen können?
Was passiert, wenn man eine politische Persönlichkeit eine beleidigende Geste machen lässt? Wenn sie Aussagen in den Mund gelegt bekommt, die sie so nie getan hat. Damit kann man ganz schön Schindluder treiben. Oder eine Staatskrise auslösen.
Wir haben es dann mit dem Missbrauch von Künstlicher Intelligenz zu tun: den sogenannten Deepfakes. Die ethischen, rechtlichen und vor allem moralischen Konsequenzen sind gewaltig.

Definition

Die Definition von Deepfake ist nicht ganz eindeutig. Meist sind mit dem Begriff realistisch wirkende Medieninhalte gemeint, die durch die Techniken der KI abgeändert und damit verfälscht worden sind.
Der Begriff selbst ist ein Kofferwort. Auf der einen Seite steht Deep, von Deep Learning, also dem automatisierten Lernen künstlicher neuronaler Netzwerke.
Auf der anderen Seite steht Fake. Die absichtliche Täuschung und Verfälschung von Inhalten.
Man fasst unter diesem Oberbegriff alle Varianten dieser Technik zusammen, ganz egal ob nur die Stimme geändert wurde, die Mimik, Gestik oder die ganze Person per Computer generiert worden ist. Man unterscheidet dann auch nicht nach dem Zweck. Ob nun zum

Spaß ein Video manipuliert wurde, ob Fehlinformationen verbreitet werden sollen oder ob der Ruf einer Person nachhaltig geschädigt werden soll. Wir haben es dann immer mit einem Deepfake zu tun.

Varianten

Es gibt mittlerweile auch unterschiedliche Varianten, die je nach Einsatz der Technik nicht ganz einfach zu unterscheiden sind. Am häufigsten begegnet man

- synthetisch erstellten Avataren
- dem Face Swapping
- und dem Face Reenactment

Es gibt mittlerweile Anbieter, die aus einem einfachen Eingabetext eine künstlich generierte Person erstellen können, zum Beispiel einen KI-Weihnachtsmann.
Ich könnte einen KI-Weihnachtsmann dazu bringen, den Rest dieses Lernvideos zu sprechen.
Doch häufiger begegnet uns das Face Swapping.
Die Face Swap Methode sieht man oft auf Social Media. Die Technik wird auch für Videofilter verwendet. Darunter versteht man das Morphen eines Gesichtes auf das einer anderen Person. Man kann mal ein Hündchen sein oder eben Tom Cruise. Damit die Täuschung möglichst gut funktioniert, nimmt man eine Person, die der Zielperson sehr ähnlich sieht, dreht neues Videomaterial.

Quelle [2, 5]

Ausgetauscht wird dann nur das Gesicht. Die passende Mimik errechnet der Computer anhand der gefilmten Vorlage. Diese Manipulationen sind dann oft schnell zu entlarven, wenn der Kontext zu absurd erscheint.

Beim Face Reenactment ist das Ganze dann schon schwieriger.

Quelle [6]

Hier nimmt man das originale Videomaterial, eine Archivaufnahme oder sogar einen abgefangenen Videostream. Eine zweite Person wird aufgezeichnet und der Computer rechnet die Mimik und Gestik dieser zweiten Person auf das originale Videomaterial. Daraus entsteht eine täuschend echte Fälschung. Wird dann noch dazu die Stimme angepasst, auch das ist technisch möglich, kann man Politiker*innen alles sagen lassen.

Aber wie funktioniert das?

Eine dafür vortrainierte KI erhält von der Person, die gefälscht werden soll, aber auch von der Person, die die neuen Bewegungen macht, ausreichend Trainingsmaterial in Form von Bildern oder Videos.

Quelle [7]

Erst dann ist sie dazu fähig, im Videomaterial selbst die gefälschten Bewegungen einzubauen. Populäre Verfahren erreichen das durch die Erzeugung eines 3D-Modells des

Gesichts der Zielperson. Im neu generierten Video trägt die Zielperson dann die täuschend echten Gesichtsausdrücke des Manipulators.
Die benötigte Datenmenge für das Training ist enorm. Kein Wunder also, dass bisher vor allem Personen des öffentlichen Lebens wie Schauspieler*innen oder Politiker*innen Ziel der Manipulationen gewesen sind.

Die Manipulation von Gesichtern in Videos ist jetzt kein neues Verfahren. Das gefälschte Archivmaterial mit John F. Kennedy ist aus dem Jahr 1997. Seitdem wird an der Technik weiter geforscht. Entwickelt von Wissenschaftler*innen weltweit, wurde die Technik über wissenschaftliche Papier publiziert. Der Code ist allen zugänglich. Das heißt auch, die Deep Fakes werden immer besser.
Mit genug Trainingsmaterial und einer bewussten Absicht für Fehlinformationen kann man, mit einem leistungsstarken Rechner, Deepfakes erzeugen und damit gehörig Schindluder treiben.

Ethische und rechtliche Grenzen

Die ethischen und rechtlichen Folgen von gefälschtem Videomaterial sind bereits heute bekannt. Und man muss keine Person des öffentlichen Lebens sein, um Opfer dieser Manipulationen zu werden.
Auch Journalist*innen werden regelmäßig Opfer von viralen Verleumdungskampagnen, wenn den anonymen Trollen die Berichterstattung nicht gefällt. Ihre Gesichter werden in Pornos gemorpht, um sie so zu verunglimpfen.

Quelle [8, 9]

Wie die vielen Einzelfälle zeigen, müssen sich die Opfer mehrere Jahre mit diesen visuellen Angriffen herumschlagen. Rechtlich gegen Deepfakes vorzugehen, ist immer noch sehr schwer.
Das Bundesamt für Sicherheit in der Informationstechnik beschäftigt sich ebenfalls mit Missbrauchsszenarien der Deepfake Technologie.

Quelle [10]

Neben den Verleumdungsfällen, die bis zum Rufmord gehen können, werden noch weitere Bedrohungsszenarien vorgestellt.
Biometrische Systeme können ausgetrickst werden. Als Schwachstelle werden Videoidentifikationsverfahren eingeschätzt.

Quelle [11]

Mittels Deepfake kann man sich als jemand anderes ausgeben und erhält Zugriff auf die Konten. Eine neue Form von Phishing-Angriffen.
Der Finanzmarkt ist aber nicht nur deswegen besorgt.

Quelle [12]

Was wäre, wenn ein Deepfake einer berühmten Person in einem Meeting Geldtransaktionen befiehlt, Kaufanfragen von Firmen stellt oder leichtfertige Äußerungen zur eigenen Firma tätigt?
Nicht auszudenken, was da auf dem Aktienmarkt los wäre.
Besonders gefährlich wird es, wenn mittels Deepfake-Verfahren glaubwürdige Desinformationskampagnen gestartet werden.

Quelle [13]

Im Rahmen von Fake News haben wir in den letzten Jahren gerade im Bereich der Bildmanipulation schon einige solcher Kampagnen erlebt. Wie schwierig wird es dann erst, wenn solche manipulierten Videos von Schlüsselpersonen kursieren.
Können wir Archivaufnahmen noch trauen, wenn sie beliebig veränderbar sind? Mittlerweile können sogar Fotos zum Leben erweckt werden.
Deepfakes sorgen dafür, dass wir es in Zukunft mit einer Vielzahl von Szenarien des Missbrauchs zu tun haben werden. Von gezielten Desinformationskampagnen bis zum Rufmord ist alles dabei.

Aber was kann man dagegen tun?

Die Technik und die stetige Verbesserung der verwendeten KI-Systeme sind nicht aufzuhalten. Hier hilft zunächst die Prävention durch Aufklärung. Die Medienkompetenz muss im Zuge dieser KI-Verfahren geschult werden.
Darüber hinaus experimentiert man gerade mit kryptografischen Verfahren. Das Ziel dabei ist, die Medieninhalte an eine vertrauenswürdige Quelle zu binden. Quasi ein Siegel für Authentizität. Aktuelle Entwicklungen testen bereits digitale Signaturen, die schon beim Aufnahmeprozess sicherstellen, dass dieses Video nicht manipuliert wurde.

Quelle [14]

Auch gesetzlich prüft man Regeln, um die Verbreitung von Deepfakes zu behindern. Die EU-Kommission fordert eine Art Regulierung, sodass Videos, die mit Deepfake-Technologie erstellt werden, auch als solche gekennzeichnet werden müssten.

Wie kann man Deepfakes erkennen?

Die trainierten Netze werden beim Erstellen von Deepfakes immer besser. Man kann jedoch heute (noch) diese Fälschungen durch ein paar Merkmale erkennen.

Quelle [1]

Dazu gehören Grafikartefakte. Diese findet man häufig am Rand des Gesichtsfelds, am Übergang vom Originalgesicht und der darauf gelegten Fälschung. Gerade wenn ein Kopf sich ins Profil dreht, kommt es zu Ausbrüchen, die deutlich zu erkennen sind.
Auch die Lichteinstellungen, die Sprachausgabe und unnatürliche Bewegungen können ein Hinweis auf eine Manipulation sein.

Deepfakes werden immer besser
Doch die Deepfakes werden immer besser. Bei gleichem Aufnahme Setting, also auch gleichem Lichtwinkel, bei ausreichend viel Trainingsmaterial von Politiker*innen oder anderen Personen des öffentlichen Lebens, sind die Manipulationen kaum noch mit dem eigenen Auge zu entdecken.
Dafür gibt es die Medienforensik. Um bewusste Fälschungen schnell zu enttarnen, ist die Weiterentwicklung der automatischen Detektion wichtig.
Diese automatisierten Systeme werden auch wieder mit großen Datenmengen trainiert. Diese wiederum werden verwendet, um die Deepfake KIs zu verbessern.

Take-Home Message
Jetzt heißt es:
Deep Learning gegen Deep Learning.
Ein Wettlauf der selbstlernenden Netzwerke. Der schnellere Rechner gewinnt.

Quellen

Quelle [1] Zsolnai-Fehér, K. (2020, April 15). *Sure, DeepFake Detectors Exist—But Can They Be Fooled?* [Video]. https://www.youtube.com/watch?v=hFZlxpJPI5w

Quelle [2] DeepTomCruise (2021, Dezember 27). *Tom on TikTok* [TikTok]. https://www.tiktok.com/@deeptomcruise/video/7046380612780952878

Quelle [3] Bregler, C., Covell, M., & Slaney, M. (1997). Video Rewrite: Driving visual speech with audio. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '97*, 353–360. https://doi.org/10.1145/258734.258880

Quelle [4] Peretti, J., & Sosa, J. (2018, April 17). *You Won't Believe What Obama Says In This Video!* [Video]. https://www.youtube.com/watch?v=cQ54GDm1eL0

Quelle [5] IamJesseRichards. (2021, September 29). *Which mask should I wear today? #deepfake*. TikTok. https://www.tiktok.com/@iamjesserichards/video/7013471575252929797

Quelle [6] Quelle Baker, H., & Capestany, C. (Regisseure). (2018, September 27). *It's Getting Harder to Spot a Deep Fake Video* [Video]. https://youtu.be/gLoI9hAX9dw

Quelle [7] GolemDE. (2019, November 21). *Eigene Deepfakes mit DeepFaceLab—Tutorial* [Video]. https://youtu.be/gfajGnx17XE

Quelle [8] White, J. (2022, April 19). *Inside the disturbing rise of 'deepfake' porn*. Dazed. https://www.dazeddigital.com/science-tech/article/55926/1/inside-the-disturbing-rise-of-deepfake-porn

Quelle [9] Ayyub, R. (2018). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. HuffPost UK.

Quelle [10] *Deepfakes—Gefahren und Gegenmaßnahmen*. (o. J.). Bundesamt für Sicherheit in der Informationstechnik. Abgerufen 17. Oktober 2022, von https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes.html?nn=1009560

Quelle [11] Tobias Fischer. (2021, April 9). Deep Fakes erleichtern Betrug. Börsen-Zeitung. https://www.boersen-zeitung.de/deep-fakes-erleichtern-betrug-e5a43b80-9933-11eb-bf5d-0b6bc58d3fae

Quelle [12] Tobias Fischer. (2021, April 10). Via Deep Fake kann jeder Chef spielen. Börsen-Zeitung. https://www.boersen-zeitung.de/banken-finanzen/via-deep-fake-kann-jeder-chef-spielen-152bdd28-7c46-11eb-9b2e-701f2e479d6b

Quelle [13] Jap San (Regisseur). (2021, März 22). 10 Most Historical Images brought back to LIFE with MyHeritage Deep Nostalgia. https://www.youtube.com/watch?v=N0lHnSHLsac

Quelle [14] European Parliament. Directorate General for Parliamentary Research Services. (2021). *Tackling deepfakes in European policy.* Publications Office. https://data.europa.eu/doi/10.2861/325063

Weiterführendes Material

Gaur, L. (Hrsg.). (2023). *Deepfakes. Creation, Detection, and Impact* (First edition). CRC Press, Taylor & Francis Group.

Pawelec, M. (2021). *Deepfakes: Technikfolgen und Regulierungsfragen aus ethischer und sozialwissenschaftlicher Perspektive* (1. Auflage). Nomos.

Disclaimer

Transkript zu dem Video „Woche 8 Ethik: Deepfake – der Missbrauch von KI", Dr. Jacqueline Klusik-Eckert.

die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.