BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG
BEIM DEUTSCHEN BUNDESTAG

Deepfakes – Manipulation von Filmsequenzen

Themenkurzprofil Nr. 25 | Marc Bovenschulte | Mai 2019

Die Geschichte der Medienmanipulation ist vermutlich so alt wie die Medien selbst. Durch die zunehmende Technisierung von Medien und Kommunikation wurden die Inhalte zum Teil zunächst glaubwürdiger (z.B. durch Beweisfotos), zugleich aber zunehmend auch Gegenstand von technischer Manipulation und Fehlinformation. Während die fotorealistische Manipulation von Standbildern ein in der Öffentlichkeit gar vermuteter oder erwarteter Standard ist – die Bildbearbeitungssoftware „Photoshop“ für das Glätten von Fotos ist Teil der Alltagssprache – stellt die Erzeugung von fiktiven, aber täuschend echten Filmsequenzen eine neue Qualität hochtechnisierter Manipulation dar; es ist ein weiterer Angriff auf die „Ich glaube nur, was ich sehe“-Überzeugung.

Die heutige Massenkommunikation erfolgt besonders bei Jugendlichen zu großen Teilen in sozialen Medien und in Form von rasch konsumierten Filmsequenzen. Daher stellen die mittels Systemen künstlicher Intelligenz (KI) erzeugten und als Deepfakes bezeichneten fiktiven Medieninhalte perspektivisch besondere Herausforderungen an Glaubwürdigkeit und Vertrauenswürdigkeit medial vermittelter Informations- und Kommunikationsinhalte und letztlich an grundlegende Diskurse und Prozesse in einer offenen und demokratischen Gesellschaft. Dies gilt auch für die jeweiligen Vertreter und Verantwortlichen der unterschiedlichen privaten oder öffentlichen Medienformate.

Hintergrund und Entwicklung

Definition und Anwendung zur Manipulation von Filmsequenzen

Das Wort Deepfake ist eine Wortkombination aus Deep Learning und Fake und beschreibt die Technik der digitalen Manipulation von Ton-, Bild- und Videomaterialien mithilfe von Deep Learning, einem Verfahren des maschinellen Lernens, das in Systemen mit KI eingesetzt wird. Zentrales Merkmal ist dabei die (foto)realistische Erzeugung fiktiver Medieninhalte oder die Manipulation bereits existierender Filmsequenzen (Chesney/Citron 2018). Diese Medienbearbeitungstechnik erlaubt es, digitale Inhalte synthetisch zu produzieren, welche eine Äußerung oder eine Aktion einer Person realistisch darstellen, ohne dass diese tatsächlich stattgefunden hat. Deepfakes, auch synthetische Medien genannt (Gregory 2018a), sind eng mit dem Konzept von Fake News verbunden – das teilweise bereits in der TAB-Untersuchung „Algorithmen und digitalen Medien und deren Einfluss auf Meinungsbildung“ (Laufzeit: 2017–2019) erforscht wird – und stellen insoweit eine neue Variante der Verbreitung von falschen oder irreführenden Informationen mit der Absicht, einer Person, einer Organisation oder einer Institution zu schaden, dar (Sängerlaub 2017). Deepfakes fügen sich ein in die lange Reihe der medialen Manipulationen zum Zweck der Falsch- oder Desinformation. Gepaart mit dem Willen, Falschmeldungen in die Welt zu setzen, ist es mit Deepfakes möglich, auf sehr überzeugende Weise den Eindruck zu erwecken, bestimmte Situationen hätten sich in der gezeigten Form ereignet (Schmid/Al-Youssef 2018).

Anwendungsfälle von Deep Fakes

Das Phänomen echt wirkender, digital erzeugter/manipulierter Filmsequenzen erlangte im Jahr 2017 größere Bekanntheit, als manipulierte Pornofilme auftauchten, die den Anschein erweckten, bekannte Popstars würden in ih-

nen mitwirken. Die betreffenden Personen wurden mithilfe einer speziellen KI-Software anhand entsprechender Vorlagen in die Filme eingefügt. Auch wenn die Qualität noch nicht sehr hoch war, ist es nur eine Frage der Zeit und des technischen Fortschritts, bis kein Unterschied zur Realität mehr feststellbar ist. Bei Verfahren des maschinellen Lernens hat sich zudem gezeigt, dass die Güte ihres Ergebnisses in großem Maße von der Qualität und der Menge der für das Lernen verwendeten Trainingsdaten abhängt. Für die menschliche Stimme wurde der Schritt zur täuschend echten Imitation mittels KI bereits vollzogen – etwa von der Software des Start-ups Lyrebird (https://lyredird.ai). Bei der digitalen Bearbeitung von Medieninhalten kommen (Sängerlaub 2017) verschiedene Formen und Zielsetzungen der Manipulation zur Anwendung:

- Offensichtliche Manipulation von Filmsequenzen, oft zum Zweck der Satire, aber auch als bewusst diffamierende, herabwürdigende oder aggressive Inszenierung.
- Schneiden und De- bzw. Rekontextualisierung von Film- und/oder Tonsequenzen und damit Erzeugung einer verdichteten, veränderten oder auch nicht mit dem Ursprung übereinstimmenden „Faktenlage“ (z.B. digitale Vergrößerung von Menschenmengen bei politischen Ereignissen; Neuordnung von Fragen und Antworten, Überspringen von maßgeblichen Sequenzen, Verschneiden von Fragmenten zu Sätzen mit neuem Inhalt), um Inhalte aus dem Kontext zu reißen, unzulässige Verkürzungen und Zuspitzungen vorzunehmen oder schlicht den Sinn zu verdrehen.
- Retusche und digitale Nachbearbeitung und damit das nachträgliche Entfernen, Einfügen und Vertauschen von Medieninhalten (im Regelfall einzelne Objekte/Personen in Fotos). Ebenso zählt dazu prinzipiell auch das „Glätten“ von Aufnahmen zum Beispiel durch verschiedene Filter und Effekte. In allen Fällen sollen „unliebsame“ Inhalte entfernt/überdeckt und erwünschte Inhalte hinzugefügt/betont werden.

Die Retusche von Filmen ist erst seit Kurzem möglich

Die Retusche von Fotos kann gewissermaßen als Vorläufer von Deepfakes angesehen werden. Im Vergleich zu Bewegtbildern ist die fotorealistische Manipulation von Standbildern mittels (kommerzieller) Werkzeuge wie Photoshop (https://www.adobe.com/de/products/Photoshopfamily.html) längst Gang und Gäbe – sei es im Kontext kommerzieller Tätigkeiten (Nachbearbeitung von Modefotografien etc.) oder zur bewussten Manipulation. Inzwischen können vielfach schon mit Smartphones Fotos mit Methoden der KI retuschiert werden und künftig sind hier weitere Fortschritte zu erwarten (Gharbi et al. 2017). Während das klassische Retuschieren von Fotos schon seit einigen Jahren durch entsprechende Software weitgehend ersetzt wurde und diese heute auch für Amateure ohne Aufwand zugänglich ist, war die fotorealistische Manipulation von Bewegtbildern bisher professionellen (Film-)Studios und Spezialisten für Visual Effects vorbehalten (Gregory 2018b), da ein erheblicher technischer Aufwand betrieben werden musste: „Im Gegensatz zu digitalen Bildern war die Videobearbeitung eine zeitaufwendige und mühsame Aufgabe, da es keine ausgeklügelten Bearbeitungstools wie Photoshop gab, aber eine große Anzahl von Bearbeitungsschritten für ein Video erforderlich ist – z.B. erfordert ein 20-Sekunden-Video mit 25 Bildern pro Sekunde eine Bearbeitung von 500 Bildern. Daher waren sehr realistische gefälschte Videos selten, und die meisten lassen sich anhand einiger auffälliger visueller Artefakte relativ leicht identifizieren.“ (eigene Übersetzung nach Li et al. 2018, S. 1)

KI als notwendige Voraussetzung für die täuschend echte Manipulation

Durch die Verfügbarkeit von Deep-Learning-Methoden hat sich die Situation nun verändert und der manuelle Aufwand reduziert sich dadurch auf ein Minimum. In sogenannten erzeugenden gegnerischen Netzwerken (Generative Adversarial Networks [GAN]) werden zwei neuronale Netze miteinander kombiniert, von denen das eine den gestaltenden Teil des Algorithmus (z.B. die Erzeugung von fiktiven Bildern) übernimmt und das andere den bewertenden Teil (z.B. die Einschätzung der Echtheit des Bildes). Auf diese Weise ist eine vollautomatisierte Manipulation von Bilddaten möglich. Durch den Lerneffekt aus der Bewertung erzeugt das gestaltende Netzwerk immer

bessere Vorschläge, sodass das Training schließlich zu einem täuschend echten Ergebnis führt (Gregory 2018b). Allerdings ist ein solches Training technisch anspruchsvoll, zeitaufwendig und abhängig von den zur Verfügung stehenden Trainingsdaten. In dem Fachbeitrag „Deep Video Portraits", in dem eine Methode zur Generierung von Videos auf Grundlage eines Austauschs der Gesichter einer Ausgangs- und einer Zielperson beschrieben wird, heißt es dazu, dass das Training des Netzwerks 10 Stunden für eine Zielvideoauflösung von 256 x 256 Pixel und 42 Stunden für 512 x 512 Pixel. dauert (Kim et al. 2018). Eine derartige Auflösung entspricht noch nicht der HD-Qualität (1.920 x 1.080 Pixel) aktueller TV-Geräte, ist für Internetvideos jedoch ohne Zweifel ausreichend.

Die Realisierung hochwertiger Deepfakes wird immer einfacher

Ausgehend von den mittlerweile frei verfügbaren Softwaretools zur Erzeugung von Deepfakes (die unter anderem auf der Open-Source-Software-Bibliothek „TensorFlow" von Google aufsetzen) durch Privatpersonen ist zu erwarten, dass aufgrund des technischen Fortschritts (steigende Rechenleistung und deren Verfügbarkeit, optimierte Algorithmen und Verfahren etc.) in absehbarer Zeit auch vermehrt hochauflösende, fotorealistische synthetische Medien/Deepfakes in Umlauf kommen. Das zuvor genannte Trainingsbeispiel wurde auf einem handelsüblichen Computer auf Basis eines Intel Xeon E5-2637 Prozessors mit 3.5 GHz und 16 GB RAM sowie einer NVIDIA GeForce GTX Titan Xp Grafikkarte mit 12 GB RAM durchgeführt (Kim et al. 2018). Während Privatpersonen in ihrer überwiegenden Mehrheit noch den programmtechnischen, zeitlichen und damit unter Umständen auch finanziellen Aufwand zur Erzeugung hochwertiger Deepfakes scheuen dürften, ist anzunehmen, dass dies für Organisationen oder staatliche Institutionen (etwa Geheimdienste und Propagandaeinrichtungen) bereits heute schon nicht mehr gilt.

Wettlauf zwischen Entwicklung und Detektion von Deepfakes

Demnach steht ein Wettlauf zwischen Verfahren zur Erzeugung von Deepfakes und deren Entlarvung mit hoher Wahrscheinlichkeit unmittelbar bevor. Ein aktuell sehr erfolgversprechender Ansatz zur Detektion von mittels GAN erzeugten Deepfakes wird wie folgt zusammengefasst: „Die neuen Entwicklungen in tiefen generativen Netzwerken haben die Qualität und Effizienz bei der Erstellung realistisch aussehender gefälschter Gesichtsvideos deutlich verbessert. In dieser Arbeit beschreiben wir eine neue Methode, um gefälschte Gesichtsvideos, die mit neuronalen Netzwerken erzeugt wurden, aufzudecken. Unsere Methode basiert auf der Erkennung von Augenzwinkern in den Videos, da dies ein physiologisches Signal ist, das in den synthetischen Videos nicht gut dargestellt werden kann. Unsere Methode wurde mithilfe von Benchmarks auf Basis von Datensätzen zu Augenzwinkerbewegungen getestet und zeigt auch eine vielversprechende Leistungsfähigkeit bei der Erkennung von Videos, die mit Deepfake erzeugt wurden." (eigene Übersetzung nach Li et al. 2018)

Gesellschaftliche und politische Relevanz

Deepfakes schwächen die Glaubwürdigkeit von Medieninhalten

Es ist zu befürchten, dass durch die Verbreitung von Deepfakes die damit verbundene abermalige Verwischung der Grenze zwischen Original und Manipulation sowohl zu einer weiteren Erosion des Vertrauens in einzelne Medien bzw. Medienformate als auch zu einer Erosion der grundlegenden gesellschaftlichen Vertrauensbasis führt (Bettilyon 2018), da keineswegs wahr sein muss, was doch „mit den eigenen Augen" gesehen wurde – ein Phänomen, das für Fotos schon länger gilt. Mit Deepfakes scheint es prinzipiell möglich, mittels einfach zugänglicher technologischer Verfahren (unter Ausnutzung gesellschaftlicher Grundstimmungen), Skandale zu provozieren und auch Streitigkeiten oder gar ernste Konflikte auszulösen, wenn selbst „Qualitätsmedien" derartige Inhalte fälschlicherweise verbreiten würden, da ihre Kontrollmechanismen – journalistische wie technische – versagen. Fake News in Form von Fake Videos erhalten eine neue, besonders leicht zu verbreitende und konsumierende sowie gleichzeitig besonders überzeugungskräftige Substanz. Zugleich kann mit dem bloßen Verdacht bzw. Vorwurf, bei einem Film (z.B. Beweisvideo) handele es sich um Deepfake, dessen Glaubwürdigkeit infrage gestellt werden – was Spezialisten womöglich trennscharf unterscheiden können, kann angesichts der Flut von Videos auf Plattformen wie Youtube bei eher flüchtiger Betrachtung zu voreiligen oder falschen Schlüssen führen. Gerade der

schnelllebige Medienkonsum in sozialen Medien sowie über Plattformen (mpfs 2018) und die bisweilen lauffeuerartige, virale Verbreitung von Videos und Informationssplittern erschweren eine solche fundierte Analyse.

Deepfakes können das Vertrauen in die Demokratie untergraben

Damit könnte der Effekt von Deepfakes auf zwei unterschiedlichen Ebenen auch unterschiedlich wirken. Während gezielte Deepfakes insbesondere die Integrität von Personen genauso wie Institutionen infrage stellen, beträfe das steigende Misstrauen in Medienhalte die Gesellschaft und Demokratie als solche: „Die Verbreitung von Deepfakes wird jenes Vertrauen untergraben, das für ein effektives Funktionieren der Demokratie notwendig ist, und zwar aus zwei Gründen. Erstens, und vor allem, wird der Marktplatz der Ideen mit einer besonders gefährlichen Form der Falschheit gefüllt. Zweitens, und subtiler, kann die Öffentlichkeit eher bereit sein, wahre, aber unbequeme Fakten zu bezweifeln. Kognitive Verzerrungen fördern bereits heute den Widerstand gegen solche Fakten, und das Bewusstsein für allgegenwärtige Deepfakes kann diese Tendenz verstärken und eine gute Ausrede bieten, um unerwünschte Beweise zu ignorieren. Insbesondere wenn gefälschte Videos weit verbreitet sind, kann die Öffentlichkeit Schwierigkeiten haben zu glauben, was ihre Augen (oder Ohren) ihnen sagen – selbst wenn die Informationen ganz real sind.“ (eigene Übersetzung nach Chesney/Citron 2018)

Mögliche vertiefte Bearbeitung des Themas

Das Phänomen der Deepfakes stellt eine neue Entwicklungsstufe der Manipulation und Verbreitung medialer Inhalte dar, für die die prinzipiell breite Verfügbarkeit hochentwickelter Technologien eine wichtige Rolle spielt. Vertiefende Untersuchungsfragen beziehen sich daher auf den gegenwärtigen Stand der Technik bei der Manipulation von Video- und Audioaufzeichnungen, zu erwartende Entwicklungsschritte in den nächsten Jahren und die Frage der Qualität sowie Unterscheidbarkeit von echten und gefälschten Filmen heute und in Zukunft. Diese Betrachtung muss insbesondere beinhalten, ob es heute und zukünftig zuverlässige Verfahren gibt, Deepfakes zu detektieren.

In diesem Zusammenhang ist zu prüfen, welche gesellschaftlichen, wirtschaftlichen und politischen Kontexte besonders durch Deepfakes betroffen sind (z.B. Strafverfolgung), welche medialen Nutzungsformen Deepfakes zu einer größeren oder geringeren Wirkung verhelfen und welche technischen und sonstigen Schutz- und Abwehrmaßnahmen hier jeweils geeignet sind. Dies gilt insbesondere für die Auswirkungen auf die politische Kommunikation und öffentliche Meinungsbildung (Schmid/Al-Youssef 2018), aber auch für die Arbeit und Rolle von Medien. Weiterhin ist zu untersuchen, welchen Beitrag eine Medienforensik beim Umgang mit

dem Phänomen und zu seiner Eindämmung oder Abwehr leisten kann und welche rechtlichen Maßnahmen und Regulierungen erforderlich bzw. zielführend sind.

Der (internationale) Sachstand zum Thema wäre z.B. in einem Thesenpapier synoptisch zusammenzufassen, zu dessen Diskussion Sachverständige zu einer öffentlichen Fachveranstaltung im Bundestag eingeladen werden. Die Ergebnisse dieser Fachveranstaltung können sodann auf Grundlage des überarbeiteten Thesenpapiers als Kurzstudie veröffentlicht werden. Allerdings stellen Deepfakes – bis auf eine überschaubare Anzahl von Ausnahmen – bislang noch kein Massenphänomen dar, sondern vor allem eine zukünftige Herausforderung, sodass dieser Umstand mittels einer stärker antizipierenden Sicht (in Form von Szenarien o.Ä.) berücksichtigt werden muss. Dies gilt auch für mögliche positive Potenziale etwa in Bildungskontexten (z.B. das Auflebenlassen von Zeitzeugen).

Literaturverzeichnis

- Bettilyon, T. E. (2018): Deep Fakes and The Future of Propaganda. Deep Fakes pose a massive threat to society and democracy, don’t get caught off guard. Medium, 11.10.2018, https://medium.com/@TebbaVonMathenstien/deep-fakes-and-the-future-of-propaganda-18a257026571 (11.2.2019)
- Chesney, R.; Citron, D. (2018): Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy? LAWFARE, 21.2.2018, https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy (12.2.2019)
- Gharbi, M.; Chen, J.; Barron, J. T.; Hasinoff, S.W.; Durand, F. (2017): Deep Bilateral Learning for Real-Time Image Enhancement. In: ACM Transactions on Graphics 36(4), S. 118:1–118:12
- Gregory, S. (2018a): Deepfakes and Synthetic Media: What should we fear? What can we do? WITNESS, 07/2018, https://blog.witness.org/2018/07/deepfakes/ (13.2.2019)
- Gregory, S. (2018b): Summary of Discussions and Next Step Recommendations from “Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions

Discovery Convening". WITNESS, 11.7.2018, https://docs.google.com/document/d/1oOvc9NtkI2m9ZlWAUlHK-1cgMkNgpsE39ld3nDGvu-N8/edit# (2.5.2019)
- Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. (2018): Deep Video Portraits. In: ACM Transactions on Graphics 37(4), S. 163:1–163:14
- Li, Y.; Chang, M.-C.; Lyu, S. (2018): In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. https://arxiv.org/pdf/1806.02877 (11.2.2019)
- mpfs (Medienpädagogischer Forschungsverbund Südwest) (2018): JIM-Studie 2018. Jugend, Information, Medien. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger (Autoren: Feierabend, S.; Rathgeb, T.; Reutter, T.). Stuttgart, https://www.mpfs.de/fileadmin/files/Studien/JIM/2018/Studie/JIM_2018_Gesamt.pdf
- Sängerlaub, A. (2017): Deutschland vor der Bundestagswahl: Überall Fake News?! Stiftung Neue Verantwortung, https://www.stiftung-nv.de/sites/default/files/fakenews.pdf (13.2.2019)
- Schmid, F.; Al-Youssef, M. (2018): Deepfake-News täuschen alle Sinne – und könnten Kriege auslösen. Der STANDARD, 13.5.2018, https://derstandard.at/2000074430944/Deepfake-News-taeuschen-alle-Sinne-und-koennten-Kriege-ausloesen (13.2.2019)

Das Horizon-Scanning ist Teil des methodischen Spektrums der Technikfolgenabschätzung im TAB.

Horizon SCANNING

Mittels Horizon-Scanning werden wissenschaftlich-technische Trends und sozio-ökonomische Entwicklungen in frühen Entwicklungsstadien beobachtet und in den Kontext gesellschaftlicher Debatten eingeordnet. So sollen Innovationssignale möglichst früh erfasst und ihre technologischen, ökonomischen, ökologischen, sozialen und politischen Veränderungspotenziale beschrieben werden. Ziel des Horizon-Scannings ist es, einen Beitrag zur forschungs- und innovationspolitischen Orientierung und Meinungsbildung des Ausschusses für Bildung, Forschung und Technikfolgenabschätzung zu leisten.

In der praktischen Realisierung wird das Horizon-Scanning als Kombination softwaregestützter Such- und Analyseschritte und eines expertenbasierten Validierungs- und Bewertungsprozesses durchgeführt.

Herausgeber: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB)
Gestaltung und Redaktion: VDI/VDE Innovation + Technik GmbH
Bildnachweise: © kanpisut/AdobeStock (S. 1), Monster Ztudio/AdobeStock (S. 2), Tatyana Gladskih/AdobeStock (S. 3), Tatyana Gladskih/AdobeStock (S. 3), Geber86/iStock (S. 4)
Stand: Mai 2019
ISSN-Internet: 2629-2874