Deepfakes - wenn wir unseren Augen und Ohren nicht mehr trauen können: Medienmanipulationen im Konflikt - Herausforderungen und Bewältigungsstrategien

Kleemann, Aldo

SWP

SWP-Aktuell

NR. 43 JUNI 2023

Deepfakes – Wenn wir unseren Augen und Ohren nicht mehr trauen können

Medienmanipulationen im Konflikt: Herausforderungen und Bewältigungsstrategien

Aldo Kleemann

Täuschung und Medienmanipulation sind seit jeher fester Bestandteil der Kriegskommunikation. Nie zuvor aber war es derart einfach, qualitativ hochwertige Fälschungen von Ton-, Bild- und Videoaufzeichnungen zu erstellen. Die menschliche Neigung, emotional auf diese Medien zu reagieren, eröffnet deren Produzenten eine völlig neue Dimension des Missbrauchs. Mit einem Kapitulationsaufruf von Präsident Selenskyj, der umgehend als Deepfake entlarvt wurde, liegt der erste Versuch eines Einsatzes der neuen Technologie in einem bewaffneten Konflikt vor. Derartige Fälschungen werden immer besser, die Erkennung immer aufwendiger und ein Ende dieser Entwicklung ist nicht absehbar. Ein Verbot von Deepfakes ist aussichtslos. Es ist deshalb Zeit, sich mit den aktuellen und potentiellen Anwendungsfällen und mit möglichen Gegenstrategien auseinanderzusetzen.

Dass die Wahrheit zu den ersten Opfern des Krieges gehört, ist altbekannt und der Einsatz von Propaganda ist ein bekanntes Phänomen in Konflikten. Zu den gängigen Mitteln zur Verbreitung von Desinformation zählen (soziale) Medien, politische Organisationen, Kulturvereine, Stiftungen und Denkfabriken. Auch wenn die Attribution von Desinformation bisweilen schwierig ist, gehören Russland und China zweifellos zu den bedeutendsten Akteuren in diesem Bereich. Deren Werkzeugkasten wird mit Hilfe von generativer Künstlicher Intelligenz (KI) seit einigen Jahren um ein weiteres Tool ergänzt: Deepfakes — täuschend echt wirkende, künstlich erstellte oder veränderte Foto-, Video- oder Sprachaufzeichnungen. Viele Deepfakes der ersten Generation ließen sich aufgrund typischer Bildfehler oder blecherner Stimmen noch recht gut entlarven. Hochwertige Fälschungen waren selten, und so attestierte die Bundesregierung Deepfakes 2020 zu Recht noch eine »geringe praktische Relevanz« und sah sich nicht gezwungen, eine spezifische Reaktionsstrategie zu erarbeiten.

Heute sind einfach nutzbare KI-Tools, mit denen sich hochwertige Fälschungen produzieren lassen, frei zugänglich. Deepfakes sind ein Massenprodukt und keine Seltenheit mehr. Dazu haben in den vergangenen Jahren vor allem drei Entwicklungen beigetragen: die kontinuierliche Verbesserung der KI, die stetige Zunahme

Stiftung Wissenschaft und Politik
Deutsches Institut für Internationale Politik und Sicherheit

verfügbarer Rechenleistung und der offene Zugang zu immer mehr Daten, mit denen die KI trainiert werden kann. Ein Ende und die volle Tragweite dieses Prozesses sind aktuell kaum absehbar.

Wie werden Deepfakes erzeugt

Im Gegensatz zu Cheapfakes, bei denen bestehende Aufnahmen manuell oder digital zusammengeschnitten, verlangsamt oder beschleunigt werden, ermöglicht die Einbindung von KI eine automatisierte Neuschöpfung oder Veränderung von Medienprodukten. So können beispielsweise Gesichter oder Sprache lippensynchron ausgetauscht, Gesten und Mimik verändert oder sogar ganze Reden erfunden und Persönlichkeiten in den Mund gelegt werden. Dazu werden in einem GAN (Generative Adversarial Network bzw. generierendes gegnerisches Netzwerk) zwei neuronale Netzwerke kombiniert und anhand vorhandener Bild-, Video- und Sprachaufzeichnungen trainiert. Insbesondere bei öffentlichen Personen stehen die dafür notwendigen Daten oft umfangreich zur Verfügung. Das anschließende »deep learning« der neuronalen Netze ist so tiefgehend und die Ergebnisse sind so realistisch, dass der heute umgangssprachliche Begriff Deepfake auf diesen intensiven Prozess zurückgeht. Im GAN findet ein Wechselspiel zwischen zwei Komponenten statt. Während der gestaltende Teil (Generator) fiktive Bilder oder Stimmen erzeugt, übernimmt der andere Teil (Diskriminator) deren Bewertung hinsichtlich der Echtheit des gegebenen Trainingsdatensatzes. Ziel ist es, dass der Generator Medien produziert, die möglichst nicht vom Trainingsdatensatz zu unterscheiden sind. Dieser Prozess kann unter anderem durch Anpassung des Datensatzes und der Gewichtung der Auswahlkriterien oder durch die Ergänzung des Diskriminators um echte Menschen stetig verbessert werden. Es ist davon auszugehen, dass die Erkennbarkeit von Deepfakes mit einer stetigen Verbesserung der genannten Parameter drastisch abnimmt.

Aktuelle Beispiele von Deepfakes

Synthetische Videos von Barack Obama und Angela Merkel zeigen, was mit den entsprechenden Trainingsdaten möglich ist. Während die Ersteller des ersten Videos einen Präsidenten Obama zeigen, der Präsident Trump beleidigt, lassen die Macher des zweiten die Bundeskanzlerin eine Rede über das Verhalten der Bundesbürger in Coronazeiten in Versform halten. Beide Videos wurden produziert, um auf die Gefahren von Deepfakes hinzuweisen, und sind als solche gekennzeichnet. Aktuell erscheinen vor allem auf Twitter immer wieder realistisch wirkende Deepfake-Bilder von bekannten Persönlichkeiten, ohne dass diese immer als solche kenntlich gemacht werden. Darunter Papst Franziskus, der einem Rapper gleich in Daunenjacke posiert, Donald Trump, wie er verhaftet wird, wie er Präsident Putin küsst oder die chinesische Flagge umarmt und küsst. Nahezu täglich kommen weitere Beispiele hinzu; manche zur Unterhaltung, manche zur Warnung und manche zur Täuschung. Am 22. Mai 2023 wurde über Twitter die Meldung einer Explosion am Pentagon verbreitet. Der Tweet erweckte den Eindruck einer offiziellen Meldung der Nachrichtenagentur Bloomberg. Begleitet war er von einem Bild, das schwarzen Rauch über dem Pentagon zeigte — ein Deepfake, der schnell erkannt wurde und dennoch ausreichte, um den US-Aktienindex S&P 500 kurzzeitig um rund 30 Punkte absacken zu lassen.

Für die Erstellung solcher Bilder braucht es lediglich eine KI-Anwendung wie Midjourney oder Stable Diffusion und eine möglichst konkrete Beschreibung des zu erstellenden Bildes.

Evolution von Desinformationskampagnen

Die Nutzung von generativer KI zur Produktion von Deepfakes verändert den Einsatz von derlei Fälschungen in Desinformationskampagnen in drei Hinsichten auf grundlegende Weise:

- **Quantität** — marktverfügbare Apps ermöglichen eine massenhafte, schnelle und kostengünstige Anfertigung von Deepfakes. Das erlaubt es neben Staaten auch ressourcenarmen Gruppierungen und Individuen, eigene Desinformationskampagnen im großen Maßstab durchzuführen.
- **Qualität** — Deepfakes werden qualitativ immer besser und wirken natürlicher, wodurch sie schwerer zu erkennen sind und an Glaubwürdigkeit und Überzeugungskraft gewinnen.
- **Qualifikation** — Während die Erstellung von Deepfakes nahezu keinerlei Qualifikation voraussetzt, wird die zu ihrer Erkennung erforderliche Expertise immer umfangreicher.

Diese Entwicklungen haben das Potential, die Reichweite und Wirksamkeit von Desinformation im 21. Jahrhundert signifikant zu erhöhen.

Einsatzmöglichkeiten von Deepfakes in Konflikten

Am 15. März 2022 kam es zum ersten nennenswerten Versuch, einen Deepfake in einem bewaffneten Konflikt zu instrumentalisieren. Nachdem die Homepage des TV-Senders Ukraine 24 gehackt wurde, erschien dort ein Video von Präsident Selenskyj, in dem dieser erklärte: »Es gibt kein Morgen mehr. Zumindest nicht für mich. Jetzt treffe ich noch eine schwierige Entscheidung: Mich von euch zu verabschieden. Ich rate [euch] die Waffen niederzulegen und zu euren Familien zurückzukehren. In diesem Krieg lohnt es sich nicht zu sterben.« Die Ukraine war auf einen solchen Deepfake-Angriff vorbereitet. Innerhalb von Minuten wurde ein echtes Antwort-Video des Präsidenten aufgenommen und über soziale Medien verbreitet. Die schlechte Qualität des Deepfakes, die schnelle Aufklärung und Erstellung einer eigenen Videobotschaft sowie die Möglichkeit, diese über eine weitgehend stabile Internetverbindung zu verteilen, haben erheblich dazu beigetragen, dass der gefälschte Kapitulationsaufruf keine Wirkung entfalten konnte. Diese Rahmenbedingungen wird man jedoch nicht in jedem zukünftigen Konflikt vorfinden. Bei anhaltender Weiterentwicklung der Deepfake-Technologie sind die Schwellen zur Produktion und zum Einsatz von Deepfakes zudem weniger technischer, sondern allein kreativer Art.

Lähmung — Deepfakes könnten in Form gefälschter Beweismittel zur Lähmung oder Spaltung von Verbündeten eingesetzt werden. Ein solcher Ansatz wurde im Vorfeld der Invasion der Ukraine debattiert. So äußerten US-Sicherheitsexperten die Vermutung, Russland plane zur Begründung des Angriffs auf die Ukraine die Anfertigung gefälschter Videobeweise für ukrainische Kriegsverbrechen an russischen Bevölkerungsgruppen. Derartige Videobeweise wären geeignet gewesen, in europäischen Staaten eine Diskussion über die Zulässigkeit einer russischen Grenzüberschreitung zum Schutz russischer Minderheiten auszulösen und eine unmittelbare Reaktion zugunsten der Ukraine zu verhindern. Im konkreten Fall wurde nicht der Einsatz von Deepfakes vermutet, sondern eine traditionelle Fälschung mit Requisiten und Schauspielern. Deepfakes könnten die Produktion solcher Szenen in Zukunft erleichtern. Die Erstellung oder Veränderung von Augenzeugenberichten oder von vermeintlich stimmauthentischen Mitschnitten von völkerrechtswidrigen Befehlserteilungen ließen sich schon heute generieren.

Mobilisierung — Deepfakes könnten auch dazu eingesetzt werden, um Bevölkerungsgruppen gegen Sicherheitskräfte zu mobilisieren. Dazu böten sich bereits vorhandene ethnische, kulturelle, soziale oder religiöse Bruchlinien in und zwischen Gesellschaften an. So könnte beispielsweise der Missbrauch religiöser Symbole vorgetäuscht werden, indem Fotos oder Videos von Schändungen erstellt oder Augenzeugenberichte gefälscht werden. Welches Mobilisierungspotential dem tatsächlichen oder vermeintlich missbräuchlichen Umgang mit religiösen Symbolen innewohnt, haben die Unruhen im Zuge des Karikaturenstreits 2005 oder der

Koranverbrennungen durch US-Streitkräfte 2012 in Afghanistan deutlich gemacht.

Zersetzung — Deepfakes ließen sich einsetzen, um Angst und Unsicherheit zu schüren. Gefälschte Kapitulationsaufrufe, herablassende Äußerungen über eigene Gefallene oder das Hinterfragen von Sinn und Zweck der militärischen Operation durch politische und militärische Führungsfiguren wären geeignet, die Streitkräfte zu demoralisieren. Desgleichen könnten Fälschungen von Grausamkeiten der eigenen Soldaten gegen die Zivilbevölkerung verwendet werden, um die Unterstützung der Bevölkerung für die Streitkräfte zu untergraben. Überdies ließen sich massenhaft sehr anschauliche Belege für die Schrecken des Krieges in Bild und Ton kreieren, um die Mobilisierung der Bevölkerung zu verhindern und Desertion zu begünstigen.

Handlungsempfehlungen

Deepfakes werden bleiben. Der Anreiz, sich zum eigenen Narrativ passende, überzeugungskräftige Medieninhalte schnell und kostengünstig erschaffen zu können, ist schlichtweg zu groß. Das zeigt sich bereits heute, außerhalb bewaffneter Konflikte, im demokratischen Diskurs. Sowohl in Deutschland als auch in den USA greifen Parteien und deren Unterstützer in der innenpolitischen Auseinandersetzung bereits auf Deepfakes zurück, um ihre Botschaften zu verstärken. Hinzu kommt, dass die zugrundeliegende KI-Technologie neben den aufgezählten negativen auch eine ganze Reihe positiver Anwendungsmöglichkeiten bietet.

Eine »silver bullet«, also eine einfache und universell einsetzbare Wunderwaffe gegen Deepfakes, wird es nicht geben. Die heutigen Einschätzungen der Möglichkeiten und Grenzen generativer KI basieren naturgemäß auf einer Momentaufnahme. Die Entwicklungsdynamik in diesem Technologiesektor hat selbst Experten und Expertinnen wiederholt überrascht. Zudem ist unklar, welche Fähigkeiten KI-Modelle haben, an denen derzeit unter privater und staatlicher Regie gearbeitet wird, und welchen Restriktionen sie unterliegen.

Viele Lösungsansätze sind daher entweder sehr spezifisch und nur auf Einzelfälle zugeschnitten oder müssen, um mit der Dynamik Schritt zu halten, holzschnittartig angelegt werden und bedürfen absehbar einer kontinuierlichen Anpassung. Notwendig ist eine Mischung aus präventiven und reaktiven Maßnahmen, um die Wirkung von Deepfakes einzudämmen.

Präventive Ansätze und ihre Grenzen

Präventive Ansätze zielen darauf ab, die Hürden für den Einsatz von Deepfakes anzuheben und deren Wirkungsmöglichkeiten von vornherein einzuschränken.

Akteure reduzieren/kontrollieren — Die Erstellung von Deepfakes erfordert spezielle Software und Hardware. Der Zugriff auf diese Ressourcen ist ein möglicher Ansatzpunkt für regulatorische Maßnahmen zur Reduktion und Kontrolle derjenigen Akteure, die in der Lage sind, einen Deepfake zu kreieren. In diesem Zusammenhang werden Exportrestriktionen für Hardwarekomponenten und Beschränkungen für den Zugriff auf Rechenleistung, Trainingsdaten oder fertige KI-Modelle diskutiert.

Ein Beispiel dafür ist die seit Oktober 2022 von den USA eingeführte Reglementierung des Exports von Halbleitern und anderen für den Bau von Supercomputern benötigten Elementen nach China. Eine solche hardwareseitige Beschränkung kann dazu beitragen, das Wachstum der Rechenleistung als Grundlage für generative KI-Modelle zu verlangsamen. Allerdings ist ein erheblicher Regelungsbedarf vorhanden, da nicht nur der direkte Export, sondern auch die indirekte Belieferung über Drittstaaten berücksichtigt werden muss, um eine solche Restriktion wirksam umzusetzen.

Verfügt ein Staat nicht über eigene Computer, ist der Rückgriff auf Cloud-Computing eine einfache Möglichkeit, Exportbeschränkungen auf Hardware-Lieferungen

zu umgehen. Daher wird mitunter vorgeschlagen, den Zugang zu Cloud-Rechenleistung zu begrenzen. Die Umsetzung einer solchen Regulierung ist in der Praxis schwierig. Einerseits müssten ihr nahezu alle weltweiten Cloud-Anbieter unterliegen, andererseits ist es schwer festzustellen, ob angemietete Rechenleistung für eine Klimasimulation oder das Training einer KI genutzt wird.

Ähnlich schwierig ist die Durchsetzung einer Beschränkung des Zugriffs auf Trainingsdaten wie Bild- und Videoaufzeichnungen. Zwar hat der Umfang der Trainingsdaten einen wesentlichen Einfluss auf die Leistungsfähigkeit der KI und seine Limitierung ist daher prinzipiell geeignet, die Anzahl der Akteure, die eine potente generative KI-Anwendung trainieren könnten, zu begrenzen. Aber auch hierfür bedürfte es der Einigkeit aller Beteiligten. Zudem ist es fraglich, ob sich eine solche Regelung für im Internet frei verfügbare Daten überhaupt wirksam implementieren ließe.

Ist eine KI fertig trainiert, entscheiden die Entwickler, wie das Modell genutzt werden kann und wer Zugang dazu erhält. Hieraus ergeben sich einige Optionen der Zugangskontrolle, die tatsächlich effektiv umgesetzt werden können. Sie greifen allerdings nur, sofern ein Großteil der Anbieter mitmacht und solange es keine Open-Source-Alternativen zu diesen KI-Modellen gibt.

Kennzeichnungspflicht — Eine Kennzeichnungspflicht durch die Endnutzer, wie sie die EU derzeit in Artikel 52 der KI-Verordnung vorsieht, ist nicht geeignet, die Zahl der Deepfakes zu reduzieren. Besser wäre eine softwareseitige Kennzeichnungspflicht. Diese hätte zur Folge, dass die gängigen, frei verfügbaren KI-Anwendungen in Europa nur noch erkennbare Deepfakes produzieren würden. Zwar ließen sich solche Kennzeichnungen entfernen, aber die dafür notwendige Expertise begrenzt den Kreis derjenigen, die einen Deepfake ohne Kennzeichnung lancieren könnten.

Sensibilisierung — Das Wissen um Deepfakes und deren Einsatzmöglichkeiten kann dazu dienen, einen kritischeren Umgang mit audiovisuellen Medien zu erlernen. Mit Blick auf Konflikte und Krisen ist dieses Wissen besonders beim Führungspersonal in der Politik und in den Behörden und Organisationen mit Sicherheitsaufgaben zu fördern. Allerdings dürfen derartige Ansätze nicht davon ausgehen, dass Deepfakes weiterhin mit bloßem Auge und ohne technische Hilfsmittel erkennbar sind.

Vertrauenswürdige Inhalte stärken — Einige Überlegungen richten sich nicht direkt darauf, Deepfakes zu erkennen, sondern Transparenz zu schaffen und so die Verbreitung von vertrauenswürdigen Ton- und Bildaufnahmen zu vereinfachen. Zu diesen Ansätzen zählt beispielsweise die Content-Authenticity-Initiative. Die beteiligten Unternehmen, darunter die BBC, Nikon, Reuters und Adobe, versuchen plattformübergreifende Industriestandards zu schaffen, die eine sichere Aussage über die Herkunft digitaler Inhalte erlauben. Ziel ist es, den Aufnahmen fälschungssichere Identitäts- und Verlaufsdaten anzuhängen, damit die Urheberschaft und jegliche Veränderungen an den Dateien dauerhaft nachvollziehbar sind. Auf diese Weise schafft der Standard Transparenz im Verbreitungsprozess, aber die Aussagekraft ist ausschließlich auf Herkunft und nachträgliche Modifikation beschränkt. Denn dass die Aufnahme an sich eine authentische Abbildung der Realität ist, kann nicht garantiert werden.

Einsatzmöglichkeiten untersuchen — Um dem Einsatz von Deepfakes in Zeiten von Krisen und Konflikten effektiv begegnen zu können, müssen Sicherheitsbehörden die Anwendungsmöglichkeiten der Technologie auch selbst untersuchen. In den USA wird einerseits vor den Gefahren von Deepfakes für die Demokratie gewarnt, gleichzeitig aber intensiv durch das Special Operations Command sondiert, wie sich die Technik militärisch verwenden lässt. Eine solche kontinuierliche Auseinandersetzung könnte in Deutschland, der EU und der Nato in bereits vorhandenen Strukturen stattfinden:

- Die seit 2019 unter Federführung des BMI arbeitende interministerielle Arbeitsgruppe zu hybriden Bedrohungen mit der dazugehörigen Expertengruppe für Desinformation ist ein geeignetes Format, um die Erfahrungen aus unterschiedlichen Ressorts zusammenzubringen.
- Im Geschäftsbereich des BMVg beobachtet das Zentrum Operative Kommunikation den Informationsraum und untersucht bereits heute die Wirkung von Propaganda auf die Streitkräfte.
- Der Austausch mit EU- und Nato-Partnern könnte über das in Helsinki ansässige Zentrum gegen hybride Bedrohungen (EU) und das in Riga gegründete Zentrum für strategische Kommunikation (Nato) erfolgen.

Reaktive Ansätze und ihre Grenzen

Reaktive Ansätze zielen darauf ab, die Wirkung eines bereits veröffentlichten Deepfakes zu reduzieren. In einer Zeit, in der die Verbreitung von Informationen nicht mehr in Tagen, sondern in Minuten gemessen wird, bedarf es der Fähigkeit zur raschen Erkennung und schnellen Antwort auf einen Deepfake.

Technische Erkennung — Die Vielzahl an Manipulationsmöglichkeiten macht es unwahrscheinlich, dass in absehbarer Zeit eine automatisierte Erkennung im Sinne einer »One-fits-all«-Lösung zur Verfügung steht. Überdies ist der wirtschaftliche Anreiz, immer bessere Deepfakes zu erstellen, derzeit deutlich höher als der Anreiz, an Techniken zu ihrer Entlarvung zu arbeiten. Dem ist von staatlicher Seite durch gezielte Förderung medienforensischer Expertise entgegenzuwirken. Die Bandbreite der Erkennungsansätze ist enorm und reicht beispielsweise von der individuellen Erfassung der Mimik und des Sprachrhythmus hochrangiger Führungspersönlichkeiten bis hin zur Erhebung von Stromnetzschwankungen, um Ort und Zeitpunkt einer Aufzeichnung zu verifizieren oder die verwendeten Geräte zu identifizieren. Um einem potentiellen Angreifer den Einsatz eines überzeugenden Deepfakes zu erschweren, ist es vor allem entscheidend, dass die Aufklärungsansätze variiert und teilweise auch geheim gehalten werden. Andernfalls wird der Diskriminator beständig an die bekannten Erkennungsverfahren angepasst, um diese gezielt zu umgehen.

Reaktionsstrategie — Eine wirksame Reaktionsstrategie umfasst viele der bereits genannten Punkte: eine grundsätzliche Sensibilität für das Thema, eine kontinuierliche Auseinandersetzung mit Deepfakes und — damit verbunden — ein Mediamonitoring sowie die Möglichkeit einer schnellen technischen Erkennung und Bewertung potentieller Fälschungen. Anschließend braucht es eingeübte Abläufe: innerhalb der Regierung, zwischen den Ressorts aber auch mit den Partnern in der EU und der Nato.

Dass die Ukraine, den Deepfake der angeblichen Selenskyj-Rede so schnell aufklären konnte, lag zum einen daran, dass der Präsident im März 2022 eine der am intensivsten beobachteten Personen in den Medien war; zum anderen auch daran, dass die ukrainischen Behörden den Einsatz von Deepfakes antizipiert hatten. Neben der Ukraine kämpfen auch andere Staaten mit anhaltender Desinformation. In Taiwan, immer wieder das Ziel chinesischer Nachrichtenmanipulation, sind die Ministerien angehalten, eigenständig innerhalb von 60 Minuten auf eine Veröffentlichung von Falschinformationen zu reagieren — eine Zeitspanne, an der sich auch Deutschland in Anbetracht der Verbreitungsgeschwindigkeit von Desinformation orientieren sollte. Das setzt eine bisher unbekannte Anpassungsfähigkeit und Reaktionsgeschwindigkeit staatlicher Institutionen voraus. Von Institutionen also, deren Entwicklungszyklen bisher eher in Jahren, wenn nicht gar Jahrzehnten gemessen wurden, nicht in Minuten und Stunden.

Oberstleutnant i.G. Aldo Kleemann ist Gastwissenschaftler in der Forschungsgruppe Sicherheitspolitik.

SWP
Stiftung Wissenschaft und Politik
Deutsches Institut für Internationale Politik und Sicherheit

Ludwigkirchplatz 3—4
10719 Berlin
Telefon +49 30 880 07-0
Fax +49 30 880 07-100
www.swp-berlin.org
swp@swp-berlin.org

ISSN (Print) 1611-6364
ISSN (Online) 2747-5018
DOI: 10.18449/2023A43