COUNTER EXTREMISM PROJECT

KONRAD ADENAUER STIFTUNG

Deepfakes

Eine Bedrohung für Demokratie und Gesellschaft

Prof. Dr. Hany Farid
und Dr. Hans-Jakob Schindler

www.kas.de

Deepfakes

Eine Bedrohung für Demokratie und Gesellschaft

Prof. Dr. Hany Farid und Dr. Hans-Jakob Schindler

Auf einen Blick

Diese Studie ist das Ergebnis einer Kooperation der Konrad-Adenauer-Stiftung mit dem Counter Extremism Projekt (CEP). Die Autoren, Prof. Dr. Hany Farid und Dr. Hans-Jakob Schindler, beschäftigen sich mit dem destruktiven Potenzial sogenannter Deepfakes. Das sind mit Hilfe künstlicher Intelligenz veränderte Fotos und Videos, die unter anderem für kriminelle und politische Manipulationszwecke eingesetzt werden.

- Zwar ist das Phänomen der Medienmanipulation nicht neu. Da die Hürden für die Nutzung der Deepfake-Technologie und deren digitale Verbreitung aber erheblich gesunken sind, kann mittlerweile von einer „Demokratisierung" der Technologie gesprochen werden.

- In Verbindung mit Desinformationskampagnen über Social Media können Deepfake-Videos disruptiv wirken und eine echte Gefahr für die Demokratie sein.

- Über Social-Media-Plattformen lassen sich Desinformationskampagnen kostengünstig weltweit verbreiten. Diese Plattformen haben unterschiedliche Standards und operieren unreguliert, was den Kampf gegen Fake News zusätzlich erschwert.

- Deutschland hat noch nicht die vollständige disruptive Kraft von Desinformationskampagnen zu spüren bekommen – im Vergleich zu anderen Ländern. Noch ist Zeit gegenzusteuern.

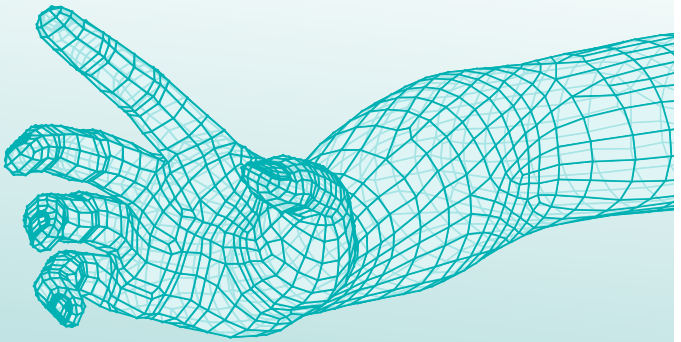
- Der Kampf gegen Deepfakes muss auf verschiedenen Ebenen geführt werden. Insgesamt wird eine Kombination rechtlicher, informationstechnologischer und bildungspolitischer Ansätze notwendig sein, um dem destruktiven Potenzial von Deepfakes zu begegnen.

Autoren

Prof. Dr. Hany Farid, Professor, University of California, Berkeley; Senior Advisor, Counter Extremism Project

Dr. Hans-Jakob Schindler, Senior Director, Counter Extremism Project

Inhaltsverzeichnis



1. Einführung

Deepfakes – mit künstlicher Intelligenz (KI) veränderte Videos, die zur Beeinflussung politischer Prozesse missbraucht werden – verleihen Des- und Fehlinformationskampagnen, die offene Demokratien und Gesellschaften angreifen sollen, eine neue Dimension. Diese Technik ist die jüngste Entwicklung in einer langen Reihe von Techniken zur Bildmanipulation, die mit der Verbreitung der Fotografie aufkamen. Früher war die Manipulation von Videos, die allgemein als exakte Abbildung der Wirklichkeit angesehen werden, nahezu ausschließlich eine Domäne von Hightech-Filmstudios und auf Spezialeffekte spezialisierten Firmen. Der technologische Fortschritt senkt die technischen Hürden jedoch und ermöglicht breiten Kreisen die Verwendung dieser Technik, einschließlich böswilliger staatlicher und nicht staatlicher Akteur/-innen.

In Verbindung mit der globalen Reichweite von Social-Media-Plattformen ermöglicht diese Technik Böswilligen, immer raffiniertere Kampagnen zur politischen Einflussnahme zu starten, die das Potenzial haben, weite Teile der Bevölkerung zu erreichen. Dies gilt umso mehr, als immer mehr Bundesbürger Nachrichten über soziale Medien konsumieren.[1]

Zum Glück ist Deutschland – im Gegensatz zu den USA – noch kein strategisches Ziel solcher Kampagnen geworden. Es gibt jedoch keinen Grund zu der Annahme, dass Deutschland in naher Zukunft von dieser Entwicklung verschont bleiben wird. Daher sollte eine strategische Diskussion über dieses Phänomen einsetzen.[2]

Auch wenn Deutschland vielleicht noch Zeit zum Reagieren bleibt, wird diese Studie verdeutlichen, dass die Entwicklung eines wirksamen Abwehrmechanismus gegen solche Bedrohungen Zeit benötigt und einen mehrgleisigen Ansatz umfassen sollte. Um einen Beitrag zu dieser bevorstehenden Debatte zu leisten, hat sich die Konrad-Adenauer-Stiftung zur Erarbeitung des vorliegenden Berichts mit dem Counter Extremism Project (CEP) zusammengetan.

Dieser Bericht besteht aus fünf Teilen. Nach der Einleitung wird Kapitel 2 einen kurzen historischen Überblick zur Bildmanipulation geben und veranschaulichen, dass die Veränderung von Bildern eine lange Geschichte hat und heute zu einer verbreiteten Praxis geworden ist, die eine Herausforderung für Medien, Rechtsverfahren, politische Debatten sowie Fragen der nationalen Sicherheit darstellt.

Kapitel 3 wird sich auf KI-gestützte synthetisierte Inhalte konzentrieren, die heute allgemein als Deepfakes bezeichnet werden. In diesem Abschnitt wird der aktuelle Sachstand bezüglich der Erstellung wie auch des Erkennens von Deepfakes umrissen, wobei betont wird, dass eine hochwertige Erkennung nach wie vor eine beträchtliche Herausforderung darstellt. Dieser Abschnitt enthält auch einen kurzen Ausblick auf zukünftige Mechanismen der Erstellung und Erkennung, wobei davor gewarnt wird, dass die zunehmende Demokratisierung dieser Technik ein wachsendes Risiko darstellen wird.

In Kapitel 4 werden diverse Vorfälle skizziert, in denen Deepfakes zu kriminellen Zwecken und mit der Absicht, politischen Einfluss zu nehmen, verwendet wurden. Es wird davor gewarnt, dass in Zeiten politischer Unsicherheit die bloße Existenz dieser Technik ausreicht, um möglicherweise politische Verwerfungen hervorzurufen, selbst wenn sie nicht genutzt wird.

Schließlich beschreibt der Bericht in Kapitel 5 relevante Elemente einer Strategie zur Abwehr der Gefahr durch Deepfakes. Eine solche Strategie sollte rechtliche Maßnahmen umfassen. Dazu könnten Beschränkungen des Einsatzes einer solchen Technik gehören sowie die Gewährleistung der Vertraulichkeit von Methoden, die zur Identifizierung von Deepfakes eingesetzt werden, und schließlich neue rechtliche Definitionen, die den Missbrauch einer solchen Technik für rechtswidrig erklären, wenn sie der intentionalen Beeinflussung politischer Prozesse tauglich scheint. Da Deepfakes, die zum Zwecke der politischen Beeinflussung missbraucht werden, ihr zersetzendes Potenzial nur bei einer weiten Verbreitung vollständig erreichen, müssen die Verbreitungsmechanismen – Social-Media-Plattformen – einbezogen werden. Zu den technikgestützten Elementen eines umfassenden Abwehransatzes, das den Einsatz von Deepfakes zwecks politischer Einflussnahme verhindern kann, sollten eine Zertifizierung von Originalinhalten und die Förderung der laufenden Entwicklung von Detektionstechniken gehören. Wirksame Maßnahmen müssen schließlich auch die Aufklärung der Öffentlichkeit beinhalten und z. B. versuchen, insbesondere (Online-)Medienkompetenz schon in der Schule zu stärken.

1 Laut der jüngsten Online-Studie von ARD und ZDF nutzen mehr als 90 % aller Deutschen über 14 Jahren Internetdienste, insbesondere Social-Media-Plattformen: ARD/ZDF Online-Studie 2019, http://www.ard-zdf-onlinestudie.de/files/2019/ARD-ZDF-Onlinestudie-Grafik-2019.pdf [06.05.2020]. Bereits im Jahr 2016 gaben rund 31 % aller deutschen Befragten an, Social-Media-Anwendungen für ihren Nachrichtenkonsum zu nutzen. Diese Zahl ist in den letzten vier Jahren sehr wahrscheinlich gestiegen (siehe: S. Hölig/U. Hasebrink, Nachrichtennutzung über soziale Medien im internationalen Vergleich. In: Media Perspektiven 11/2016, S. 534–548, https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2016/11-2016_Hoelig_Hasebrink.pdf [06.05.2020]).

2 N. Lossau, Deep Fake: Gefahren, Herausforderungen und Lösungswege, Konrad-Adenauer-Stiftung, Analysen & Argumente Nr. 382/2020, https://www.kas.de/documents/252038/7995358/AA382+Deep+Fake.pdf/de479a86-ee42-2a9a-e038-e18c208b93ac?version=1.0&t=1581576967612 [06.05.2020].

2. Eine kurze Geschichte der Bildmanipulation

Das Fälschen von Bildern zieht sich durch die Geschichte. Schon Stalin, Mao, Hitler, Mussolini, Castro und Breschnew ließen Fotos manipulieren, um die Geschichte umzuschreiben. Dazu bediente man sich komplizierter und zeitaufwändiger Techniken während der Entwicklung des Fotos in der Dunkelkammer. Seit Anfang der 1990er Jahre erleichtert die leistungsstarke und kostengünstige Digitaltechnik jedoch fast jedem und jeder, digitale Bilder zu verändern. Die hieraus entstehenden Fälschungen sind mitunter schwierig zu erkennen. In den vergangenen zwei Jahrzehnten hat diese Manipulation von Bildern viele verschiedene Bereiche betroffen.

Medien: Adnan Hajj war berühmt für seine eindrucksvollen Kriegsfotografien aus dem andauernden Nahostkonflikt. Am 7. August 2006 veröffentlichte die Nachrichtenagentur Reuters eines der Fotos von Hajj, das die Folgen eines israelischen Bombenangriffs auf eine libanesische Stadt zeigt (siehe unterer Teil von Abbildung 1[3]). In der Folgewoche berichteten Hunderte von Bloggern und fast alle größeren Medien, dass das Foto manipuliert worden sei; es wurde Rauch nachträglich hinzugefügt. Die Reaktionen zeigten Empörung und Wut – Hajj wurde vorgeworfen, das Bild manipuliert zu haben, um die Wirkung des israelischen Angriffs zu überhöhen. Die blamierte Agentur *Reuters* zog das Foto rasch zurück und nahm fast 1.000 von Hajj beigesteuerte Fotos aus ihren Archiven. Der Fall Hajj ist natürlich keinesfalls einzigartig. 2003 manipulierte Brian Walski, ein erfahrener Kriegsfotograf, ein Bild, das auf der Titelseite der *Los Angeles Times* zu sehen war. Nachdem sie von der Fälschung erfahren hatten, entließen deren empörte Herausgeber Walski. Die Nachrichtenmagazine *Time* und *Newsweek* wurden jeweils von ähnlichen Skandalen erschüttert, als bekannt wurde, dass Fotos auf ihren Titelseiten manipuliert worden waren. In den letzten Jahren mussten unzählige Nachrichtenagenturen auf der ganzen Welt ähnliche Erfahrungen machen.

Abbildung 1: Das ursprüngliche (oben) und das gefälschte (unten) Foto der israelischen Bombardierung einer libanesischen Stadt
Foto: Adnan Hajj/Reuters

Wissenschaft: Journalist/-innen sind nicht die einzigen, die versucht sind, Fotos zu manipulieren. 2004 veröffentlichten Professor Hwang Woo-Suk und Kolleg/-innen scheinbar bahnbrechende Fortschritte in der Stammzellenforschung. Ihre Arbeit wurde in *Science* veröffentlicht, einer der weltweit renommiertesten wissenschaftlichen Zeitschriften. Langsam kristallisierte sich heraus, dass die darin präsentierten Ergebnisse manipuliert und/oder fingiert worden waren. Nach monatelangen Kontroversen zog Hwang das *Science*-Papier zurück und räumte seinen Posten an der Universität. Ein unabhängiges Gremium, das die Vorwürfe untersuchte, stellte unter anderem fest, dass mindestens neun der elf Fotos von individuellen Stammzellkolonien, die Hwang angeblich kultiviert hatte, durch Manipulation fingiert waren. Dieser Fall zog zwar internationale Aufmerksamkeit auf sich und löste weitverbreitete Empörung aus, aber er ist keineswegs einzigartig. In einem wettbewerbsorientierten Bereich sind Wissenschaftler zunehmend versucht, ihre Ergebnisse zu übertreiben oder zu fingieren. Mike Rossner, Chefredakteur der Zeitschrift *Cell Biology*, schätzt, dass bis zu 20 Prozent der bei seiner Zeitschrift eingereichten Manuskripte mindestens eine Darstellung enthalten, die aufgrund unangemessener Bildmanipulation überarbeitet werden muss.[4]

Rechtswesen: Die Anklage wegen Kinderpornografie gegen ihren Polizeichef schockierte die Kleinstadt Wapakoneta im US-Bundesstaat Ohio. Vor Gericht argumentierte der Anwalt des Angeklagten, dass der Besitz der Bilder durch den Angeklagten nicht illegal sei, wenn der Staat die Echtheit der beschlagnahmten Fotos nicht beweisen könne. 1996 wurde die US-weite Strafgesetzgebung gegen Kinderpornografie durch das Gesetz *Child Pornography Prevention Act* (CPPA) ergänzt, das bestimmte Arten „virtueller" Pornografie aufführte. 2002 befand der Oberste Gerichtshof der Vereinigten Staaten von Amerika, dass Teile des CPPA zu weit gefasst und restriktiv waren und daher die im Ersten Verfassungszusatz festgelegten Rechte verletzten. Das Gericht entschied, dass der Besitz „virtueller" oder „computergenerierter" Bilder, die einen fiktiven Minderjährigen darstellen, nicht gegen die Verfassung verstoßen. Nach dieser Logik liegt die Beweislast dafür, dass die Bilder real und nicht computergeneriert sind, beim Staat. Angesichts der Raffinesse computergenerierter Bilder wurde ferner in mehreren Urteilen auf US-Bundes- und föderaler Ebene festgestellt, dass von Geschworenengerichten nicht verlangt werden sollte,

zwischen realen und virtuellen Bildern zu unterscheiden. Mindestens ein Bundesrichter stellte selbst die Fähigkeit von Sachverständigen, diese Entscheidung zu treffen, infrage.

Politik: „Fonda spricht bei Friedenskundgebung zu Vietnam-Veteranen" lautete die Schlagzeile, begleitet von einem Foto, das angeblich Senator John Kerry zeigte, wie er sich mit der Schauspielerin und umstrittenen Friedensaktivistin Jane Fonda eine Bühne teilt (Abbildung 2)[5]. Sowohl der Artikel als auch die Abbildung waren Fälschungen. Das Foto wurde aus zwei nicht miteinander zusammenhängenden Bildern zusammengesetzt. 2008, nur wenige Tage nachdem sie zur Kandidatin für das Stellvertreteramt des US-Präsidentschaftskandidaten John McCain gewählt worden war, kursierten im Internet gefälschte Bilder einer in Bikini gekleideten und eine Schusswaffe tragenden Sarah Palin. Die Idee, neben politischen Gegner/innen eine umstrittene Person zu platzieren, ist gewiss nicht neu. Man nimmt an, dass ein gefälschtes Foto zur Wahlniederlage von Senator Millard Tydings 1950 beigetragen hat. Das Foto von Tydings im Gespräch mit Earl Browder, einem Führer der Kommunistischen Partei der USA, sollte suggerieren, dass Tydings Sympathien für den Kommunismus

Abbildung 2: Eine Fotozusammenstellung von Senator Kerry und der Kriegsgegner-Aktivistin Jane Fonda Foto: Ken Light, Owen Franken/ The Guardian

Abbildung 3: Ein manipuliertes Foto, in das die dritte Rakete von links, deren Start gescheitert ist, nachträglich eingefügt wurde Foto: Sepahnews/ The Guardian

hegte. Während der jüngsten Vorwahlen in den USA wurde in politischen Anzeigen eine erstaunliche Anzahl manipulierter Fotos gezeigt, die Kandidaten in einem schmeichelhaften oder schädlichen Licht darstellen.

Nationale Sicherheit: Während sich 2008 die Spannungen zwischen den USA und dem Iran verstärkten, verkündete die iranische Regierung die erfolgreiche Erprobung ballistischer Raketen. Als Beweis dafür veröffentlichte sie ein Foto, das den gleichzeitigen Abschuss von vier Raketen zeigt. Kurz nach seiner weltweiten Veröffentlichung stellte sich heraus, dass das Bild manipuliert worden war. In Wahrheit waren nur drei Raketen gestartet, während die vierte, deren Start fehlschlug, digital auf dem Bild hinzugefügt worden war (Abbildung 3)[6]. Dieses Beispiel zeigt, wie wichtig der Zugriff auf authentische Nachrichtenbilder ist, und verdeutlicht die geopolitischen Schockwellen, die durch gefälschte Fotos hervorgerufen werden können.

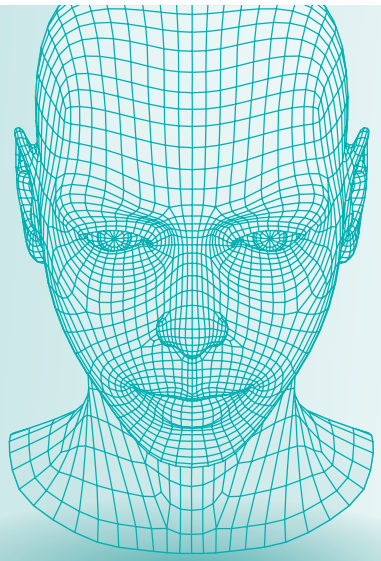
Obwohl sie in geschichtlicher Betrachtung eher eine Ausnahme darstellen, beeinflussen verfälschte Bilder in den letzten zwei Jahrzehnten zunehmend fast jeden Aspekt der Gesellschaft. Zuletzt haben uns die Fortschritte in der künstlichen Intelligenz und im maschinellen Lernen sowie der Zugang zu riesigen Datensätzen und Rechenleistungen der nächsten Revolution in der digitalen Manipulation ein großes Stück nähergebracht.

3 Foto (Original und verändert) erstellt von Adnan Hajj. Geändertes Foto ursprünglich vertrieben von Reuters.

4 H. Pearson, Image manipulation: CSI: Cell biology. In: Nature 434/ 2005, S. 952–953, https://www.nature.com/articles/434952a.pdf?proof=true&draft=collection%3Fproof%3Dtrue [06.05.2020].

5 Das Bild ist eine Kombination aus zwei Originalfotos. Das Originalfoto von Senator Kerry wurde 1971 von Ken Light aufgenommen, das Originalfoto von Jane Fonda 1972 von Owen Franken.

6 Das veränderte Bild wurde ursprünglich von der iranischen Nachrichtenorganisation Sepahnews.com veröffentlicht. Das Bild ist derzeit auf der Website nicht mehr zugänglich. Damals veröffentlichte Associated Press auch das Bild von Abbildung 3.



3. KI-synthetisierter Inhalt (alias Deepfakes)

3.1 Erzeugung von Deepfakes

Jüngste Fortschritte in den Bereichen Computergrafik, computergestützte Bilderkennung und maschinelles Lernen haben die automatische Generierung überzeugend gefälschter Audio-, Bild- und Videoinhalte erleichtert. Im Audiobereich ist nun eine äußerst realistische Audiosynthese möglich. Ein neuronales Netzwerk, das mit einer ausreichenden Anzahl relevanter Tonaufzeichnungen gespeist wird, kann lernen, Sprache auf der Grundlage der Stimme eines Benutzers oder einer Benutzerin zu synthetisieren.[7] Im Bereich statischer Bilder können nun sehr realistische Fotos von Menschen synthetisch erzeugt werden.[8] Und im Videobereich können hochrealistische Videos einer beliebigen Person erstellt werden, die praktisch alles sagt und tut, was der Erzeuger des Videos will.[9]

Abbildung 4: Eine Hillary Clinton Imitatorin (links) und ein Face-Swap-Deepfake (rechts)
Foto: Saturday Night Life/ National Broadcasting Company (NBC)

Unser Schwerpunkt liegt auf Deepfake-Videos. Solche manipulierten Videos fallen in eine von drei Kategorien: (1) Der Austausch von Gesichtern (*Face Swap*): Das Gesicht in einem Video wird automatisch durch dasjenige einer anderen Person ersetzt. Abbildung 4, die zeigt, wie das Gesicht einer Imitatorin mit dem von Hillary Clinton ausgetauscht wird, ist ein Beispiel für diese Technik.[10] Oft wird diese auch verwendet, um berühmte Schauspieler/-innen in Filmclips einzubauen, in denen sie nie aufgetreten sind, aber auch um Personen, zumeist Frauen, in Deepfakes mit pornografischem Inhalt einzufügen; (2) Lippensynchronisation (*lip sync*): Ein Originalvideo wird modifiziert, um die Muskelbewegungen im Mundbereich mit einer willkürlichen Audioaufnahme in Einklang zu bringen. Der Schauspieler und Regisseur Jordan Peele hat ein besonders überzeugendes Beispiel solch eines Deepfakes produziert, indem er ein Video des ehemaligen US-Präsidenten Barack Obama so verändert hat, dass dieser sagt: „Präsident Trump ist ein totaler und vollständiger Vollidiot."; und (3) sogenannte Puppenspieler/-innen (*puppet master*): Eine Zielperson wird von einem/-r Darsteller/-in, der/die vor einer Kamera sitzt, animiert (Kopf- und Augenbewegungen, Mimik). Die Aufnahmen dieser Bewegungen werden dann synthetisch an die Bewegungen des/-r Darsteller/in angepasst. Der/die Puppenspieler/-in bestimmt also, was seine Puppe tun und sagen soll.

Es gibt eine Fülle von Techniken zur Erstellung dieser Art von Deepfake-Videos, darunter *DeepFake FaceSwap, FSGAN*, *Neuronal Textures*, *Face2Face* und *FaceSwaps*, von denen die meisten leicht zugängliche Open-Source-Projekte sind.[11] Der meistverbreitete – aber keineswegs einzige – Ansatz zur Erstellung von Deepfake-Videos (oder -Bildern) nutzt die Möglichkeiten generativer gegnerischer Netzwerke (*generative adversarial networks*, GAN). Ein GAN besteht aus zwei Hauptkomponenten: einem Generator und einem Diskriminator. Das Ziel des Generators ist es, jedes Videobild so zu synthetisieren, dass es mit der Verteilung eines Trainingsdatensatzes übereinstimmt. Ziel des Diskriminators ist es, festzustellen, ob das synthetisierte Videobild als zum Trainingsdatensatz gehörend erkannt werden kann oder nicht. Generator und Diskriminator arbeiten iterativ, wobei der Generator schließlich lernt, ein Video – Bild für Bild – zu synthetisieren, das den Diskriminator täuscht.

Die beliebte *FaceSwap*-Software etwa verwendet ein GAN, um Face-Swap-Deepfakes zu erzeugen. Bekannte Beispiele dieses Ansatzes haben den Schauspieler Nicholas Cage in Filme eingesetzt, in denen er gar nicht mitgespielt hatte, darunter „sein" überaus unterhaltsamer Auftritt in *The Sound of Music*. Diese Technik kann zwar sehr überzeugende Fälschungen erzeugen, erfordert jedoch oft eine umfassende Menge Trainingsdaten. Die neuere Methode *FSGAN* dagegen erzeugt hochqualitative Fälschungen mit weniger Trainingsdaten.

Neural Textures ist ein generisches Bildsynthesesystem, das traditionelle Grafikdarstellung mit neueren lernfähigen Komponenten kombiniert. Dieser Rahmen kann zur Erzeugung neuer Ansichten, zur Szenenbearbeitung und zur Erstellung sogenannter lippensynchroner Deepfakes verwendet werden, die den Mund einer Person derart verändern, dass er mit einer abweichenden Audioeingabe korreliert. Dieses System kombiniert frühere Bemühungen zur Erstellung lippensynchroner Deepfakes auf individueller Grundlage.

Im Gegensatz zu diesen GAN-basierten Techniken stützen sich andere Methoden auf traditionellere Ansätze der Computergrafik, um Deepfakes zu erstellen. *Face2Face* etwa ermöglicht die Erstellung von Puppenspieler-Deepfakes, bei denen Gesichtsausdruck und Kopfbewegungen der spielenden Person auf eine andere Person (die Puppe) übertragen werden. Auf ähnliche Art generiert *FaceSwap* ein dreidimensionales

Gesichtsmodell einer Person und projiziert es auf die Aufnahme einer anderen Person. Diese Techniken ermöglichen den Benutzern und Benutzerinnen, mit handelsüblichen Kameras Modelle für Gesichtsausdrücke in Echtzeit zu erzeugen. Eine verwandte Technik, das *Puppet-Mastering*, ist in der Lage, ein fotorealistisches Avatar-GAN aufzubauen, das Mimik und Perspektive in Echtzeit auf mobilen Geräten synthetisiert.

3.2 Erkennung von Deepfakes

Die digitale Forensik hat umfangreiche Forschungsarbeiten zur Erkennung traditionell manipulierter Bilder und Videos hervorgebracht.[12] Wir beschränken uns hier auf Techniken zur Erkennung oben beschriebener Arten von Deepfake-Videos. Es gibt drei Hauptkategorien zur Erkennung von Deepfake-Videos: (1) manuelle Überprüfung, (2) *Low-Level*-Computertechniken sowie (3) *High-Level*-Computertechniken.

Ältere Deepfake-Videos wiesen offenkundige Bearbeitungsspuren wie unscharfe und verschwommene Stellen auf. Verfälschten Aufnahmen von Gesichtern fehlten typische Eigenschaften, wie etwa das Augenblinzeln.[13] Während ein kundiges Auge zwar auch aufwändige Deepfakes erkennen kann, wird die visuelle Unterscheidung echter Aufnahmen von Fälschungen immer schwieriger. Da sich die Qualität von Deepfakes stetig verbessert, bedarf die Erkennung von Deepfake-Videos automatischer und rechnergestützter Techniken.

Low-Level-Ansätze konzentrieren sich auf die Erkennung im KI-Syntheseprozess eingeführter Artefakte auf Pixelebene. Einer dieser Ansätze verwendet ein künstliches neuronales Netz (*convolutional neural network,* CNN) zur Erkennung von Artefakten auf Pixelebene, das während des Prozesses der Transposition einer Fläche auf eine andere entsteht. Ein anderer lernbasierter Ansatz trainiert ein neuronales Zwillingsnetzwerk darauf, Inkonsistenzen zwischen einem Bild und den Metadaten der Kamera (z. B. Brennweite, ISO, Blende, Belichtungszeit usw.) zu finden. Ein Bild wird dann mithilfe dieses Netzwerks authentifiziert, indem festgestellt wird, ob die einzelnen Bildpunkte mit den Metadaten konsistent sind. Das ManTra-Net (Manipulation Tracing Network)[14] ist zwar nicht notwendigerweise auf Deepfakes ausgerichtet, verwendet jedoch das *End-to-End-Training* eines CNN, um verschiedene Arten der Bildmanipulation zu erkennen und zu lokalisieren, einschließlich Spleißen, Entfernen und Klonen von Objekten (*copy-move*). Ein weiterer Ansatz hebt auf die Erkennung und Lokalisierung von Gesichtsmanipulationen ab, wobei ein Netzwerk verwendet wird, um ein Gesicht ganzheitlich als manipuliert oder nicht manipuliert einzuordnen. Ein zweites Netzwerk nutzt *Low-Level*-Merkmale auf kleinen Punkten, um festzustellen, ob eine Gesichtsregion mit dem Rest des Bilds konsistent ist; eine endgültige Aussage ergibt sich aus der Verbindung dieser beiden Beurteilungen. Andere Ansätze haben gezeigt, dass GAN-generierte Inhalte eindeutige digitale Fingerabdrücke enthalten, die identifiziert werden können und eine Aussage darüber ermöglichen, ob ein Bild GAN-generiert ist oder nicht.

Der Vorteil dieser und ähnlicher Low-Level-Ansätze ist, dass sie automatisch Artefakte und Unterschiede zwischen synthetischen und realen Inhalten ausweisen können. Nachteilig ist, dass sie sehr empfindlich auf absichtliche oder unabsichtliche Bearbeitung, einschließlich Größenänderung oder Transkodierung, sowie auf gegnerische Angriffe und auf eine Extrapolation auf neue Datensätze reagieren können. Im Gegensatz dazu sind die nachfolgend beschriebenen *High-Level*-Ansätze tendenziell widerstandsfähiger gegen diese Arten der Manipulation und wahrscheinlich robuster, wenn sie auf neuartige Datensätze angewendet werden.

High-Level-Ansätze konzentrieren sich auf semantisch bedeutsamere Merkmale. Zum Beispiel hat man in früheren Arbeiten erkannt, dass die Erstellung von Deepfakes mit Gesichtsaustausch zu Inkonsistenzen in der Kopfhaltung führt, da diese aus dem zentralen, transponierten Teil des Gesichts und dem umgebenden, ursprünglichen Kopf extrapoliert wird. Diese Inkonsistenzen wirken sich auf die 3-D-Geometrie aus und lassen sich bislang mit Synthesetechniken nur schwerlich korrigieren. Da Trainingsdatensätze oft keine Darstellungen von Personen mit geschlossenen Augen enthalten, fiel bei frühen mittels *Face Swap* erstellter Deepfakes ein ungewöhnlich seltenes Augenblinzeln auf. Bei neueren Deepfakes ist dieses Problem jedoch offenbar behoben. Eine ähnliche Technik nutzt räumliche und zeitliche physiologische Merkmale, die in echtem Videomaterial nicht konsistent aufgezeichnet wer-

den und *Face-Swap*-Deepfakes stören. In anderen Forschungsarbeiten wurden stundenlange Videoaufzeichnungen bestimmter Personen (in diesem Fall Regierungschefs aus der ganzen Welt und US-Präsidentschaftskandidaten) analysiert, um eindeutige und vorhersehbare Muster von Mimik und Kopfbewegungen zu gewinnen. Ähnliche Forschungen ergaben, dass lippensynchrone Deepfakes bei der Transposition bestimmter Töne (Phoneme) den Mund nicht präzise nachbilden.

Vorteile dieser *High-Level*-Ansätze bestehen darin, widerstandsfähiger als *Low-Level*-Ansätze gegen rechnerische Korrekturangriffe und geeigneter zu sein, eine große Bandbreite an Deepfakes zu erkennen, von *Face-Swapping* über Lippensynchronisation bis hin zu *Puppet-Mastering*. Nachteil ist der mögliche größere Aufwand bei Entwicklung und Erprobung sowie Einsatz.

Trotz der Bemühungen von Digitalforensiker/-innen, sowohl *Low-* als auch *High-Level*-Techniken zu entwickeln, existiert keine Technik, die es mit der großen Bandbreite von Deepfakes hinsichtlich Geschwindigkeit und Genauigkeit, die ein internetbasierter Einsatz ermöglicht, aufnehmen könnte.

Digitalforensik steht mehreren Herausforderungen gegenüber. Deepfakes sind ein relativ neues Phänomen. Was ihre Raffinesse betrifft, haben sie sich deutlich rascher entwickelt als erwartet. Es gibt wesentlich mehr Forscher/-innen, die an der Synthese immer realistischerer Audio-, Bild- und Videodaten arbeiten, als solche, die derlei Inhalte zu erkennen versuchen. Das heißt, Art und Qualität von Deepfakes entwickeln sich in einem unerhörten Tempo, mit dem Schritt zu halten schwierig ist. Zudem fordern Dimension und Geschwindigkeit des Internets den Einsatz wirksamer Technik beträchtlich heraus: Facebook beispielsweise verzeichnet täglich etwa eine Milliarde Uploads[15] und jede Minute werden auf YouTube etwa 500 Stunden Video hochgeladen.[16] Die schiere Menge täglich hochgeladener Informationen erschwert den Einsatz effektiver Filtertechnik spürbar.

Die auf breiter Basis einsetzbaren *Control-Capture*-Techniken können die Echtheit von Inhalten bestätigen, indem sie zum Zeitpunkt der Aufzeichnung aus jedem aufgenommenen digitalen Inhalt eine eindeutige

digitale Signatur extrahieren, sie kryptografisch signieren und dann auf einem sicheren zentralen Server oder einem verteilten, unveränderlichen Kassenbuch *(Ledger)*, wie z. B. der Blockchain, ablegen. Diese Signatur kann dann mit jeder online gefundenen Version desselben Inhalts verglichen werden, um festzustellen, ob dieser verändert wurde. Dieser Ansatz geht zwar anders mit Desinformation um als forensische Techniken – denn er meldet, was echt, und nicht, was gefälscht ist – aber die Technik ist bereits verfügbar und kann auch großflächig im Internet eingesetzt werden. Sowohl diese *Control-Capture-* als auch die klassischen forensischen Techniken sollten weiter untersucht werden.

3.3 Die Zukunft der Erzeugung und Erkennung von Deepfakes

Es bedarf derzeit einer hohen Rechenleistung, um ein langes und visuell überzeugendes Deepfake-Video zu erstellen. Diese Form der Rechenleistung ist in der Cloud zwar leicht verfügbar, hat jedoch ihre Grenzen. Die Software zur Erstellung von Deepfake-Videos ist indes leicht und frei online verfügbar. Der sich abzeichnende Trend bei der Erstellung von Deepfakes ist, dass die Qualität weiter zunimmt, während die Menge benötigter Daten und die notwendige Rechenleistung zurückgehen. Außerdem treten zunehmend kommerzielle Produkte und Websites in Erscheinung, was die Verbreitung und damit die „Demokratisierung" der Technik weiter beschleunigt. Wie weiter unten erörtert wird, bleiben zugleich die Kanäle zur Verbreitung von Deepfakes über soziale Medien leicht zugänglich. Während sich die breite Öffentlichkeit, die tagtäglich mit Falschinformationen, Komplotten und Lügen überschwemmt wird, um eine sinnvolle Einordnung der Welt um sich herum bemüht, ist zu beobachten, dass sich ein Mehrwert für Lügner (sogenannte *liar's dividend*) entwickelt.[17] Da gefälschte Informationen das Online-Biotop im Würgegriff haben, kann jeder von Nachrichten, die ihm nicht passen oder seine Weltsicht nicht bestätigen, einfach behaupten, sie seien gefälscht. Gerade diese Verbreitungs- und Einordnungsmuster stellen womöglich eine größere Bedrohung für Demokratie und Gesellschaft dar als die gefälschten Inhalte selbst.

7 A. v. d. Oord/S. Dieleman/H. Zen/K. Simonyan/O. Vinyals/A. Graves/N. Kalchbrenner/A. Senior/K. Kavukcuoglu, Wavenet: A generative model for raw audio, 2016, https://arxiv.org/abs/1609.03499 [06.05.2020].

8 T. Karras/S. Laine/T. Aila, A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, 2019, S. 4401–4410; T. Karras/S. Laine/M. Aittala/J. Hellsten/J. Lehtinen/T. Aila, Analyzing and improving the image quality of stylegan, 2019, https://arxiv.org/abs/1912.04958 [06.05.2020].

9 R. Tolosana/R. Vera-Rodriguez/J. Fierrez/A. Morales/J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, 2020, https://arxiv.org/abs/2001.00179 [06.05.2020].

10 Das Originalvideo auf der linken Seite ist ein Screenshot aus einer Episode der beliebten Comedyshow *Saturday Night Life* der National Broadcasting Company (NBC) im Jahr 2016. Das Deepfake-Bild rechts wurde zu Illustrationszwecken erstellt.

11 Siehe zum Beispiel: B. Paris/J. Donovan, Deepfakes and Cheap Fakes. The Manipulation of Audio and Visual Evidence, 2019, S. 15; https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf [27.05.2020].

12 H. Farid, Photo Forensics. Cambridge, MA/London: MIT Press. 2016.

13 Y. Li/M.-C. Chang/S. Lyu, In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In: IEEE International Workshop on Information Forensics and Security, 2018, S. 1–7, https://arxiv.org/pdf/1806.02877.pdf [06.05.2020].

14 Y. Wu/W. Abdalmageed/P. Natarajan, ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: IEEE Conference on Computer Vision and Pattern Recognition, 2019, https://openaccess.thecvf.com/content_CVPR_2019/papers/Wu_ManTra-Net_Manipulation_Tracing_Network_for_Detection_and_Localization_of_Image_CVPR_2019_paper.pdf [06.05.2020].

15 D. Noyes, The Top 20 Valuable Facebook Statistics – Updated January 2020, Zephoria, https://zephoria.com/top-15-valuable-facebook-statistics/.

16 J. Hale, More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute, Tubefilter, 7. Mai 2019, https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/ [05.06.2020].

17 R. Chesney/D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security. In: California Law Review 107/2019, S. 1753–1819, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 [06.05.2020].

4. Bedrohungen für Demokratie und Gesellschaft: Von staatlichen zu nicht staatlichen Akteur/-innen

4.1 Missbrauch von Deepfakes zu kriminellen Zwecken

Ein böswilliger Einsatz von Deepfakes stellt vonseiten sowohl staatlicher als auch nicht staatlicher Akteure eine Bedrohung dar. Vornehmliches Einsatzgebiet von Deepfake-Videos ist deren Verwendung für kriminelle Zwecke, insbesondere bei der Erstellung nicht einvernehmlicher Pornografie.[18] Dies ist nach wie vor eine Gefahr für alle Frauen und ganz besonders Prominente und Journalistinnen, die regelmäßig in der Öffentlichkeit stehen. Als Reaktion darauf haben mehrere US-Bundesstaaten vor Kurzem Gesetze verabschiedet, um den durch solche Inhalte verursachten Schaden zu minimieren. Ähnliche Gesetze werden in den USA auf Bundesebene sowie auf internationaler Ebene erwogen.

Zudem wird die Demokratisierung ausgefeilter Techniken zur Erzeugung hochrealistisch gefälschter Audio-, Bild- und Videoaufnahmen die Bekämpfung betrügerischer Fehl- und Desinformationskampagnen voraussichtlich erschweren. Ein anschauliches Beispiel hierfür ereignete sich im März 2019, als Betrüger synthetische Stimmimitationen ein-

setzten, um den Vorstandsvorsitzenden eines Unternehmens in einem Telefongespräch zu imitieren und eine Überweisung von 200.000 EUR zu erwirken.[19] Zu den weniger raffinierten Deepfakes gehört das Erstellen synthetischer Fotos nicht existenter Personen,[20] um künstliche Identitäten zum Zwecke des Betrugs und der Spionage zu erschaffen.[21] Schließlich haben Experten auch darauf verwiesen, dass Deepfakes dazu missbraucht werden können, lokale Konflikte zu verschärfen.[22]

4.2 Deepfakes zur Einflussnahme auf politische Prozesse

Schwerpunkt dieses Berichts ist der Missbrauch von Deepfakes zur Störung demokratischer Wahlen und Erzeugung gesellschaftlicher Verwerfungen. Solche Manipulationsversuche haben in den letzten zwei Jahren stetig zugenommen. Im Januar 2019 identifizierte die weltweite Bedrohungsanalyse der US-Nachrichtendienste Deepfakes als eine der größten globalen Bedrohungen:

> *Demnach werden Gegner und strategische Rivalen versuchen, mittels Deepfakes oder ähnlicher maschineller Lernverfahren generierte überzeugende – aber falsche – Bild-, Audio- und Videodateien gegen die USA und deren Verbündete und Partner einzusetzen.*[23]

Bekannte Fälle versuchter politischer Beeinflussung 2019 betrafen Deepfakes minderer Qualität. Beispielsweise scheinen bei den jüngsten britischen Parlamentswahlen Ende 2019 Fehlinformationen, darunter Deepfakes, strategisch eingesetzt worden zu sein, um die Erfolgsaussichten der Führung der Liberaldemokrat/-innen und eines prominenten Labour-Politikers bei den Wahlen zu schmälern.[24]

In den sozialen Medien kursierten zwei manipulierte Videos der aktuellen Sprecherin des US-Repräsentantenhauses Nancy Pelosi. Die Videos wurden ohne ausgefeilte Technik leicht verlangsamt, um Pelosi betrunken wirken zu lassen.[25] Obwohl die Manipulation leicht erkennbar war, wurde das Video von politischen Gegner/-innen Pelosis geteilt, selbst nachdem es als manipuliert gekennzeichnet worden war.[26] Interessanterweise weigerte sich Facebook, eine der wichtigsten Plattformen, auf denen das Video geteilt wurde, es zu löschen.[27]

Die Risiken, die sich aus der laxen Online-Überwachung selbst bei den großen Social-Media-Plattformen ergeben, wurden deutlich, als Twitter im Februar 2020 auf einen falschen Kandidaten für den US-Kongress hereinfiel. Im Dezember 2019 kündigte Twitter an, die Accounts der Kandidat/-innen für die US-Wahlen 2020[28] auf ihre Echtheit hin zu prüfen und dabei mit *Ballotpedia* zusammenzuarbeiten, einer gemeinnützigen Organisation, die eine Datenbank mit politischen Kandidat/-innen unterhält.[29] Offenbar erkannte Twitter jedoch nicht, dass ein 17-jähriger Schüler einen gefälschten Account für einen fiktiven US-Kongresskandidaten erstellt hatte, indem er ein Foto von einer Website mit einer Sammlung synthetisch erzeugter Bilder gefälschter Personen verwendete. Twitter[30] bestätigte die Echtheit des Accounts.[31]

Einen merkwürdigen Versuch offener politischer Beeinflussung stellt ein minderwertiges, für die flämische sozialistische Partei sp.a in Belgien produziertes Deepfake-Video dar, das die Partei 2018 auf ihrer Website publizierte. Darin forderte US-Präsident Donald Trump vorgeblich, dass Belgien aus dem Pariser Klimaabkommen aussteige. Die als Donald Trump posierende Figur spricht Englisch; niederländische Untertitel wurden hinzugefügt. Der einzige Teil des gefälschten Videos ohne Untertitel ist der letzte Satz, in dem der Darsteller das Filmmaterial als Fake bezeichnet.[32] Wie Kommentare auf der Facebook-Seite der Partei zeigen, erkannten viele, die das Video gesehen hatten, nicht, dass es sich um eine Fälschung handelte.[33] Offenbar beabsichtigte die Partei, die Wähler auf eine Online-Petition umzuleiten, in der die belgische Regierung aufgefordert wurde, stärkere Maßnahmen gegen den Klimawandel zu ergreifen. Da einige das Video jedoch für echt hielten, sah sich die Partei gezwungen, öffentlich zu betonen, dass es ein Witz und eine Fälschung sei. Dieser Fall veranschaulicht, welchen Einfluss Deepfakes haben können und dass deren Wirkungen von den Macher/-innen nicht immer kontrolliert werden können.[34]

Zur politischen Manipulation eingesetzte Deepfakes bergen insofern noch ein weiteres Risiko, als sie das Vertrauen der Öffentlichkeit in die politischen Institutionen untergraben. Daher kann auch politischer

Schaden entstehen, wenn keine Deepfakes im Spiel sind, wie es 2018 in Gabun der Fall war. Der Präsident Gabuns Ali Bongo erkrankte offenbar 2018 und trat mehrere Monate lang nicht öffentlich auf. Als die Regierung ein Video mit einer Neujahrsansprache von ihm veröffentlichte, wurden Vorwürfe laut, das Video sei ein Deepfake, obwohl nichts diese Behauptungen zu untermauern schien.[35] Die Vorwürfe waren offenbar Teil einer Kampagne einiger Militärs, die für einen späteren Zeitpunkt im Jahr einen Staatsstreich planten.[36] In diesem Fall reichte also die potenzielle Verfügbarkeit von Deepfake-Technik aus, um politische Verwerfungen zu verursachen.[37]

Zwar sind derzeit keine Fälle bekannt, in denen terroristische Gruppen Deepfakes verwenden, um politische Turbulenzen auszulösen, doch könnte die technische Verfügbarkeit zu Problemen bei der strafrechtlichen Verfolgung terroristischer Kämpfer/-innen führen, die aus dem Ausland zurückkehren. Einige aktuelle Fälle in Europa befassen sich mit schweren Verbrechen europäischer Kämpfer/-innen des sogenannten Islamischen Staats (IS).[38] In entsprechenden Gerichtsverfahren werden Bilder und Videomaterial als Beweismittel verwendet, deren Echtheit die Angeklagten nun glaubhaft verleugnen könnten. Die Staatsanwaltschaft stünde dann vor der neuen technischen Herausforderung, die Echtheit der Dokumente zu beweisen.

Da Deepfakes immer einfacher herzustellen sind und keine größere Rechenleistung mehr benötigen, ermöglichen sie böswilligen staatlichen Akteur/-innen schließlich die Schaffung einer zusätzlichen Distanz zwischen sich und den Deepfakes, die sie im Rahmen ihrer Kampagnen zur politischen Manipulation in Umlauf bringen können. Dadurch wird das ohnehin schon große Problem der Zuschreibung noch verschärft,[39] da ihnen das Outsourcing der Herstellung und Verbreitung von Deepfakes als Teil ihrer Fehlinformationskampagnen helfen wird, ihre Verantwortung glaubhaft zu leugnen – ein Teil der sogenannten *liar's dividend*.

In den letzten Jahren haben koordinierte Fehlinformationskampagnen böswilliger staatlicher Akteur/-innen sowohl in den USA[40] als auch in Europa[41] an Zugkraft gewonnen. Mit Unterstützung durch die Deepfake-Technik könnten solche Kampagnen ihre zersetzende Wirkung erheblich verstärken. Wie die obigen Beispiele zeigen, haben Deepfakes bei solchen Versuchen bislang keine größere Rolle gespielt. Daher bleibt noch genügend Zeit für die Entwicklung eines vielschichtigen Abwehransatzes, um den potenziellen Folgen eines Einsatzes dieser Technik im Rahmen von Kampagnen zur politischen Einflussnahme entgegenzuwirken. Im nächsten Kapitel werden rechtliche, technikbasierte und bildungspolitische Maßnahmen als denkbare Komponenten eines solchen Verteidigungssystems behandelt. Da die Entwicklung eines solchen Systems einige Zeit in Anspruch nehmen wird, sollten die ersten Schritte zu dessen Schaffung schon jetzt unternommen werden.

---

18 Laut einer Studie von Deeptrace, einem auf die Suche nach Deepfakes im Internet spezialisierten Unternehmen, betrafen 94 % (13.254 von 14.678) der von dem Unternehmen 2019 identifizierten Deepfake-Videos nicht einvernehmliche Pornografie: H. Ajder/G. Patrini/F. Cavalli/L. Cullen, The State of Deep Fakes. Landscape, Threats and Impact, Deeptrace, September 2019, S. 6, https://share.hsforms.com/1cg_h2aPnRrufZeN8HDjWPw3hq83.

19 C. Stupp, Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Scams using artificial intelligence are a new challenge for companies, Wall Street Journal, 30. August 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [06.05.2020].

20 Aus Bildern realer Personen synthetisiert.

21 Ajder/Patrini/Cavalli/Cullen, The State of Deep Fakes, S. 13.

22 T. Simonite, Forget Politics. For Now, Deepfakes Are for Bullies Wired, 4. September 2019, https://www.wired.com/story/forget-politics-deepfakes-bullies/ [06.05.2020].

23 D. R. Coats, Statement for the Record: Worldwide Threat Assessment of the Intelligence Community, 29. Januar 2019, S. 7, https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf [06.05.2020].

24 The Soufan Center, IntelBrief: The Use of Disinformation in the British Election, 13. Dezember 2019, https://thesoufancenter.org/intelbrief-the-use-of-disinformation-in-the-british-election/ [06.05.2020].

25 D. Harwell, Faked Pelosi videos, slowed to make her appear drunk, spread across social media, Washington Post, 24. März 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [06.05.2020].

26 A. Fichera, Manipulated Video Targeting Pelosi Goes Viral, FactCheck.Org, 24. Mai 2019, https://www.factcheck.org/2019/05/manipulated-video-targeting-pelosi-goes-viral/ [06.05.2020].

27 J. Waterson, Facebook refuses to delete fake Pelosi video spread by Trump supporters, Guardian, 24. Mai 2019, https://www.theguardian.com/

technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site [06.05.2020].

28 B. CoyneB. Coyne, Helping identify 2020 US election candidates on Twitter, Twitter, 12. Dezember 2019, https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html [06.05.2020].

29 Ballotpedia: About, https://ballotpedia.org/Ballotpedia:About [06.05.2020].

30 This Person Does Not Exist, https://thispersondoesnotexist.com [06.05.2020]; R. Metz, These people do not exist. Why websites are churning out fake images of people (and cats), CNN, 28. Februar 2020, https://edition.cnn.com/2019/02/28/tech/ai-fake-faces/index.html [06.05.2020].

31 D. O'SullivanD. O'Sullivan, A high school student created a fake 2020 candidate. Twitter verified it, CNN, 28. Februar 2020, https://edition.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html [06.05.2020]; Twitter hat das Konto inzwischen gesperrt.

32 J. Lytvynenko, A Belgian Political Party is Circulating a Trump Deepfake Video, BuzzFeed, 20. Mai 2018, https://www.buzzfeednews.com/article/janelytvynenko/a-belgian-political-party-just-published-a-deepfake-video [06.05.2020].

33 H. v. d. Burchard, Belgian socialist party circulates 'deep fake' Donald Trump video, Politico, 21. Mai 2018, https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/ [06.05.2020].

34 O. Schwartz, You thought fake news was bad? Deep fakes is where news goes to die, Guardian, 12. November 2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth [06.05.2020].

35 A. Breland, The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink, Mother Jones, 15. März 2019, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [06.05.2020].

36 Ajder/Patrini/Cavalli/Cullen, The State of Deep Fakes, S. 10.

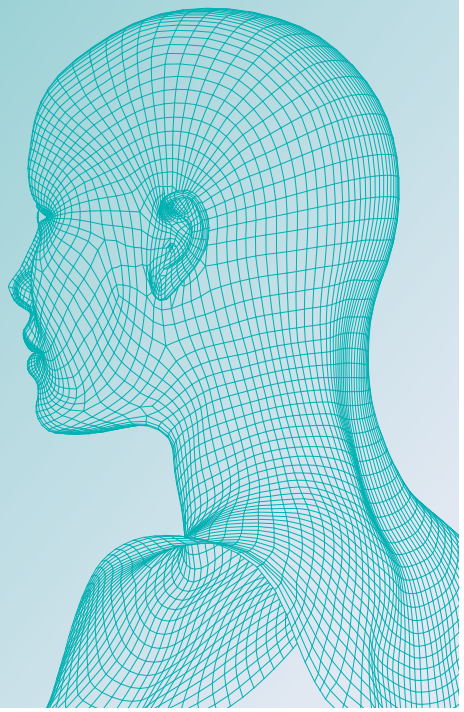
37 S. Cahlan, How misinformation helped spark an attempted coup in Gabon, Washington Post, 13. Februar 2020, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/ [06.05.2020].

38 P. Bąkowski/L. Puccio, Foreign fighters – Member State responses and EU action, European Parliamentary Research Service, März 2016, S. 8, https://www.europarl.europa.eu/EPRS/EPRS-Briefing-579080-Foreign-fighters-rev-FINAL.pdf [06.05.2020].

39 S. Bradshaw/P. N. Howard, The Global Disinformation Disorder: 2019 Global Inventory of Organised Social Media Manipulation, Working Paper 2/2019, Oxford, UK: Project on Computational Propaganda, S. 9, https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf [06.05.2020].

40 K. Roose/S. Frenkel/N. Perlroth, Facebook, Google and Twitter Struggle to Handle November's Election, New York Times, 29. März 2020, https://www.nytimes.com/2020/03/29/technology/facebook-google-twitter-november-election.html [06.05.2020].

41 Siehe zum Beispiel: L. Benková, The Rise of Russian Disinformation in Europe. In: Austria Institut für Europa- und Sicherheitspolitik, Fokus 3/2018, https://www.aies.at/download/2018/AIES-Fokus_2018-03.pdf [06.05.2020].



5. Umgang mit der Bedrohung durch Deepfakes

Deepfake-Technik ist für sich genommen selbstverständlich weder gut noch schlecht – sie kann für positive Zwecke eingesetzt werden, z. B. in der Kunst[42] oder zur Verbreitung öffentlicher Botschaften in mehreren Sprachen.[43] Zu allfälligen kriminellen Zwecken oder/und zur Beeinflussung politischer Prozesse verwendete Deepfake-Technik birgt allerdings das Potenzial, weitreichenden sozialen Schaden anzurichten.

Angesichts der Tatsachen, dass sich diese Technik weiterverbreitet, eine hohe Zahl an Nutzern erreichen kann sowie Hard- und Softwareanforderungen kein Problem mehr darstellen, sollte eine Reihe von Maßnahmen, die eine Minimierung des sozialen Schadens durch den Missbrauch dieser Technik bezwecken, erörtert werden. Dabei muss die Redefreiheit aber weiterhin gewährleistet werden und der Einsatz dieser Technik in positiven Kontexten wie Kunst und Kultur sichergestellt sein.

Vor dem Hintergrund des derzeitigen Entwicklungstempos in Technik und Verbreitung dürfte keine Einzelmaßnahme allein ausreichen, um das Problem wirksam anzugehen. Zum Beispiel dürften gesetzliche Verbote

für den kommerziellen Handel kaum wirken, da diese Technik via Internet weltweit frei verfügbar ist und Deepfake-Videos auf einer Vielzahl von Plattformen verbreitet werden.

Daher ist eine Reihe juristischer, technikbasierter und bildungspolitischer Maßnahmen zu ergreifen, die sowohl auf die Produktion als auch – was noch wichtiger ist – auf die Distribution von Deepfake-Videos abzielen.

5.1 Juristische Maßnahmen

5.1.1 Mögliche gesetzliche Beschränkungen der Technik

Hinsichtlich der Produktion von Deepfake-Videos liegt ein zentraler Ansatzpunkt darin, ein gewisses Maß an Kontrolle über die neue Technik zu behalten. Hierzu bedarf es einerseits Bemühungen, technische Hürden aufrechtzuerhalten, die nicht staatliche Akteur/-innen einschließlich Krimineller potenziell an der Anwendung dieser Technik hindern. Vor allem aber ist die Schaffung eines Rechtsrahmens bedeutsam, um böswillige Akteur/-innen nach entsprechender Zuordnung verfolgen zu können. Das hieße, den Einsatz bestimmter Arten von Deepfake-Software innerhalb einer Rechtsordnung für illegal zu erklären.

Die andere, womöglich noch wichtigere rechtliche Einschränkung betrifft den breiten Zugang zur Erkennungstechnik. Wenn diese Technik Markt und Anwender/-innen vollständig zugänglich gemacht würde, könnten böswillige Akteur/-innen ihre Produktionsmethoden rasch anpassen. Das sich hieraus ergebende technische „Wettrüsten“ ließe sich verlangsamen, indem fortschrittliche Detektionstechnik vom Markt ferngehalten wird. Daher könnten gesetzliche Schutzmaßnahmen, die die Freigabe fortschrittlicher Detektionstechnologien verhindern, eine wirksame Maßnahme sein.

Um gegen den böswilligen Einsatz von Deepfakes vorzugehen, bedarf es einer Unterscheidung zwischen Deepfakes im Zusammenhang mit kriminellen Handlungen, wie Betrug oder Herstellung nicht einvernehmlicher Pornografie, und solchen zum Zweck der Einflussnahme auf politische Prozesse. In beiden Fällen ist es bedeutsam, sowohl den Missbrauch als auch die Verbreitung zu bekämpfen – allerdings auf verschiedenen Ebenen. Um kriminelle Machenschaften ins Visier zu nehmen, sollte die missbräuchliche Verwendung von Deepfake-Technik das Hauptanliegen sein. In solchen Fällen wird ein Individuum oder eine Gruppe von Individuen geschädigt, sodass der gesellschaftliche Schaden maßgeblich von der Häufigkeit solcher Delikte abhängt.

Hinsichtlich des Missbrauchs von Deepfakes als Mittel zur politischen Beeinflussung, wobei beträchtlicher gesellschaftlicher Schaden auch durch einen einzigen Vorgang verursacht werden kann – was den Schwerpunkt dieses Berichts bildet –, ist zunächst ebenfalls der Missbrauch zu einem strafbaren Vergehen zu machen, wenngleich die Gesetzgebung angemessene und notwendige Ausnahmen vorzusehen hat, etwa für Satire, Komödie oder politische Kritik. Da im politischen Sinne manipulative Deepfakes ihr volles zersetzendes Potenzial aber nur dann erreichen, wenn sie weit verbreitet werden, ist die gezielte Bekämpfung der Verteilungsmechanismen solchen Materials ein zweiter, ebenso wichtiger Ansatz. In den folgenden Abschnitten seien daher Maßnahmen umrissen, die gegen den Missbrauch von Deepfakes zum Zwecke der Beeinflussung politischer (Wahl-)Prozesse gerichtet sind.

5.1.2 Legislative Maßnahmen gegen den Missbrauch zur Beeinflussung politischer Prozesse

Die derzeitigen Regulierungsmaßnahmen zielen auf den Missbrauch dieser Technik zur politischen Beeinflussung ab, indem neue rechtliche Kategorien geschaffen werden. Zum Beispiel verabschiedete der US-Bundesstaat Kalifornien 2019 mit dem AB 730 das erste Gesetz, das den Einsatz von Deepfakes in böswilliger Absicht bei politischen Kampagnen verbietet:

> *(...) a person, firm, association, corporation, campaign committee, or organization shall not, with actual malice, produce, distribute, publish, or broadcast campaign material that contains (1) a picture or photograph of a person or persons into which the image of a candidate*

for public office is superimposed or (2) a picture or photograph of a candidate for public office into which the image of another person or persons is superimposed.
(...)
within 60 days of an election at which a candidate for elective office will appear on the ballot, distribute, with actual malice, materially deceptive audio or visual media.[44]

AB 730 enthält eine wichtige Bestimmung, die das Recht auf die Verbreitung von Deepfakes als Teil einer Nachrichtensendung oder politischen Satire sichert:

This section does not apply to an internet website, or a regularly published newspaper, magazine, or other periodical of general circulation, including an internet or electronic publication, that routinely carries news and commentary of general interest, and that publishes materially deceptive audio or visual media prohibited by this section, if the publication clearly states that the materially deceptive audio or visual media does not accurately represent the speech or conduct of the candidate.[45]

Dieses Gesetz ist ein erster Schritt zum Schutz des politischen Diskurses vor dem böswilligen Einsatz von Deepfake-Technik. In einer Online-Umgebung wird seine Anwendung jedoch die schwierige Frage der Zuschreibung berücksichtigen müssen. Es wird sich erst mit der Zeit herausstellen, ob die anhaltende Herausforderung, ein Deepfake-Video rechtlich einwandfrei einem bestimmten Täter zuzuordnen, die wirksame Anwendung dieser neuen kalifornischen Rechtsvorschrift zum Schutz vor Deepfakes zwecks Beeinflussung politischer Prozesse in vielen Fällen verhindern wird.

Das Gesetz rief umgehend Kritik seitens derjenigen hervor, die eine unangemessene Einschränkung der Meinungsfreiheit befürchteten, insbesondere im politischen Umfeld.[46] Andere meinten, das Gesetz sei „zu schwach", um die gewünschte Wirkung zu erzielen, zumal es voraussetzt, dass das Material in böswilliger Absicht hergestellt wird.[47] Zudem gilt das Gesetz nur für die Gerichtsbarkeit Kaliforniens, bezieht also keine Akteur/-innen außerhalb des Bundesstaats ein, was einige dazu veranlasst hat, Maßnahmen auf Bundesebene zu fordern.[48]

Ein ähnlicher Gesetzentwurf wurde tatsächlich im Dezember 2019 im US-Kongress eingebracht, ist aber seitdem nicht vorangekommen.[49] Im Dezember 2019 verabschiedete der US-Kongress jedoch das Genehmigungsgesetz zur nationalen Verteidigung (*National Defense Authorization Act*, NDAA) für das Haushaltsjahr 2020.[50] Dessen Abschnitt 5709 sieht einen neuen Jahresbericht vor, der dem Kongress vom Geheimdienst-Direktor vorzulegen ist und sich auf Folgendes bezieht:

(A) the potential national security impacts of machine-manipulated media (commonly known as "deepfakes"); and

(B) the actual or potential use of machine-manipulated media by foreign governments to spread disinformation or engage in other malign activities.[51]

Gemäß Abschnitt 5709 sollte dieser Bericht auch Folgendes enthalten:

An updated identification of the counter-technologies that have been or could be developed and deployed by the United States Government, or by the private sector with Government support, to deter, detect, and attribute the use of machine-manipulated media and machine-generated text by foreign governments, foreign-government affiliates, or foreign individuals, along with an analysis of the benefits, limitations and drawbacks of such identified counter-technologies, including any emerging concerns related to privacy.[52]

Abschnitt 5724 des NDAA richtet auch einen mit fünf Millionen US-Dollar dotierten Deepfake-Award ein, „to stimulate the research, development, or commercialization of technologies to automatically detect machine-manipulated media."[53]

Somit hat der US-Kongress, obwohl auf Bundesebene kein politischer Konsens darüber besteht, wie die böswillige Verwendung von Deepfakes unter Kontrolle gebracht werden kann, deren potenziellen Einsatz im Sinne einer Einflussnahme auf politische Prozesse als Bedrohung für die nationale Sicherheit erkannt.

Diese gesetzgeberischen Bemühungen spiegeln das Entstehen und Wachsen eines spezifischen Regelwerks wider, das sich auf die Bekämpfung von Deepfakes konzentriert. Natürlich könnten die politischen Entscheidungsträger nicht nur neue Gesetze entwerfen, sondern sich auch auf bestehende, wie z. B. das Urheberrecht oder das Recht auf Öffentlichkeit (das Recht am eigenen Bild), stützen. Die bestehenden Gesetze stoßen bei diesem Thema jedoch an ihre Grenzen.[54]

Bemühungen seitens des Gesetzgebers, die sich auf den Missbrauch von Deepfakes konzentrieren, können daher nur ein Element der Reaktion sein. Politisch motivierte Deepfakes können nur dann ihr volles Störungspotenzial ausschöpfen, wenn sie eine weite Verbreitung erreichen. Daher sind neben den gesetzgeberischen, auf deren Herstellung und Missbrauch abzielenden Anstrengungen die Mechanismen zur Verbreitung dieses Materials in Betracht zu ziehen – ein zweites, ebenso wichtiges Element, um den Missbrauch dieser Technik zu verhindern.

5.1.3 Maßnahmen gegen die Verbreitung von Deepfakes

Zu Zwecken der Einflussnahme auf politische Prozesse hergestellte Deepfakes werden über Internetdienste in Umlauf gebracht. Dabei ist es wichtig, zwischen dem individuellen Austausch von Deepfakes, d. h. deren Verbreitung über persönliche Kommunikation wie E-Mail und private Nachrichten, und ihrem Austausch auf öffentlichen Plattformen, insbesondere über soziale Medien, zu unterscheiden. Regulatorische Maßnahmen, die sich auf persönliche Kommunikation konzentrieren, würden einen erheblichen Eingriff in den Bereich gesetzlich geschützter persönlicher Kommunikation erfordern und sollten daher nicht primärer Ansatzpunkt sein. Die individuelle Verbreitung dürfte trotz der Verfügbarkeit automatisierter Massen-E-Mail-Software auch langsamer vonstattengehen als die globale Massenverbreitung über Internetplattformen innerhalb von Sekundenbruchteilenerfolgen kann. Daher hat ein potenzielles Ziel eines Deepfakes-Angriffs bei der Verbreitung über persönliche Kommunikationsmittel oft mehr Zeit zu reagieren.

Die öffentliche Verbreitung über globale soziale Medien gewährleistet eine sofortige Wirkung dieses Materials. Angesichts dieses beträchtlichen den politischen Diskurs zersetzenden Potenzials müssen die Plattformanbieter/-innen, insbesondere diejenigen mit einer bedeutenden Anzahl von Nutzern, einen Beitrag zu dem zu schaffenden Abwehransatz leisten.

In den USA beschneidet der *Communication Decency Act* (CDA) von 1996 die Ansätze erheblich, Plattformanbieter/-innen gesetzlich dazu zu verpflichten, Abwehrmaßnahmen gegen den Missbrauch ihrer Dienste durch Einzelpersonen oder Organisationen einzurichten, die schädliche Inhalte veröffentlichen. Abschnitt 230 des CDA sieht vor:

> *No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.*
> *(…)*
> *No provider or user of an interactive computer service shall be held liable on account of –*
>
> *(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or*
>
> *(B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).*[55]

Das heißt, es steht Plattformanbieter/-innen frei, den Zugang zu Inhalten, die sie für unerwünscht halten, einzuschränken, zu blockieren oder zu entfernen. Sie sind jedoch nicht haftbar gegenüber Dritten, wenn diese durch das Versäumnis der Plattformanbieter/-innen, eine solche Verbreitung zu verhindern, geschädigt werden. Eine Ausnahme besteht lediglich für Inhalte im Zusammenhang mit Sexhandel. Folglich kann weder das Entfernen noch das Nicht-Entfernen von Inhalten allgemein gerichtlich verfolgt werden.

Trotz dieser Einschränkungen hat der politische Druck auf die Branche in den USA dazu geführt, dass mehrere große Plattformen neue Strategien gegen die Verbreitung von Deepfakes angekündigt haben, beispielsweise seit Ende 2019 Facebook,[56] Twitter,[57] TikTok,[58] Reddit[59] und Google/YouTube[60] entsprechende Richtlinienänderungen. Hinsichtlich ihrer Reichweite unterscheiden sich diese neuen Richtlinien jedoch erheblich voneinander.

So änderte TikTok seine Politik durch die Einführung einer sehr weit gefassten Definition, die diverse Formen von Deepfake-Material umfasst:[61]

> *Inhalte, mit denen Nutzer*innen [der] Community betrogen oder irregeführt werden sollen, gefährden das Vertrauen, auf dem unsere Community basiert. Wir dulden keine solchen Inhalte auf unserer Plattform. Dies umfasst Aktivitäten wie den Versand von Spam-Nachrichten, Vortäuschung einer anderen Identität oder Desinformationskampagnen.*
> *(...)*
> **Vortäuschung einer anderen Identität**
> *Wir dulden keine Nutzer*innen, welche die Identität anderer Personen oder Organisationen vortäuschen, um die Öffentlichkeit irrezuführen. Wenn sich Meldungen vorgetäuschter falscher Identität bestätigen, entfernen wir die an solchen Verstößen beteiligten Konten. Wir machen Ausnahmen bei Persiflage, Kommentaren oder Fan-Konten, solange die Konten andere Nutzer*innen in Bezug auf Identität oder Zweck auf TikTok nicht irreführen.*
>
> **Nicht erlaubt sind:**
> *Die irreführende Verwendung des Namens, der biografischen Angaben oder des Profilbilds einer anderen Person oder Organisation.*
>
> **Irreführende Informationen**
> *Wir gestatten keine Fehlinformationen, die unserer Community oder der Öffentlichkeit insgesamt schaden könnten. [...] Außerdem entfernen wir Inhalte, die im Rahmen von Desinformationskampagnen verbreitet werden.*
>
> **Nicht erlaubt sind:**
> *Fehlinformationen, mit denen Angst, Hass oder Vorurteile geschürt werden sollen*
> *(...)*
> *Inhalte, mit denen Mitglieder der Community in Bezug auf Wahlen oder andere zivilgesellschaftliche Prozesse irregeführt werden sollen.*[62]

Der Vorteil einer so weit gefassten Definition besteht darin, dass sie potenziell ein breites Spektrum böswilliger Aktivitäten und Formen von Deepfake-Material umfasst. Das Unternehmen definiert jedoch nicht im Detail, was es unter „Irreführung", „Täuschung" oder „Desinformationskampagnen" versteht, sodass es bei der Umsetzung seiner Politik einen liberalen Ansatz verfolgen kann. Daher wird die Qualität der internen Entscheidungsmechanismen der bestimmende Hauptfaktor sein, ob solche Plattformen den Missbrauch ihrer Dienste verhindern können.

Am anderen Ende des Spektrums möglicher Definitionen von Deepfakes hat Facebook, die derzeit größte Social-Media-Plattform, eine außergewöhnlich enge Definition gezogen, die nur die fortschrittlichsten Deepfake-Videoproduktionen abdeckt. In seinen aktualisierten Community-Richtlinien heißt es:

> **Grundgedanke dieser Richtlinie**
> *Medien, einschließlich Bilder, Audio oder Videos, können auf vielfältige Arten bearbeitet werden. In vielen Fällen sind diese Ver-*

änderungen harmlos, wie beispielsweise durch einen Filtereffekt bei einem Foto. In anderen Fällen ist die Manipulation jedoch nicht offensichtlich und könnte Nutzerinnen und Nutzer irreführen, insbesondere bei Videoinhalten. Unser Ziel ist es, diese Kategorie manipulierter Medien zu entfernen, wenn die weiter unten dargelegten Kriterien erfüllt wurden.
(...)
Folgende Inhalte sind untersagt:
Videos, die neben Anpassungen zur Erhöhung der Deutlichkeit oder Qualität so bearbeitet oder zusammengestellt wurden, dass dies für eine Durchschnittsperson nicht erkenntlich ist. Dadurch könnte eine Durchschnittsperson zu der irrigen Ansicht gelangen, dass jemand in dem Video etwas gesagt hat, das er in Wahrheit nicht gesagt hat

UND
durch künstliche Intelligenz oder maschinelles Lernen entstanden sind, einschließlich Deep Learning-Techniken (z. B. technische Deepfakes). Dabei werden Inhalte auf einem Video verschmolzen, kombiniert, ersetzt und/oder überlagert, so dass ein Video entsteht, das echt wirkt.

Diese Richtlinie betrifft keine Inhalte, bei denen es sich um eine Parodie oder Satire handelt oder die bearbeitet werden, um Wörter auszulassen, die gesagt wurden, oder die Reihenfolge gesprochener Wörter zu ändern.[63]

Mit dieser Definition decken die Richtlinien nur Deepfake-Videos ab, die mittels hochentwickelter Techniken wie KI, maschinellen Lernens oder *Deep Learning* verändert wurden. Deepfakes minderer Qualität, selbst wenn sie potenziell der Einflussnahme auf politische Prozesse dienen, sind nicht eingeschlossen, sodass Facebook deren Verbreitung gänzlich ignorieren kann.

Dieser eng gefasste Ansatz von Facebook rief umgehend Kritik hervor: Da derzeit die überwiegende Mehrheit irreführender Videos mit weniger ausgefeilter Technik produziert wird, wird diese Unternehmenspolitik dem Problem der Deepfakes zur Beeinflussung politischer Vorgänge nicht gerecht.[64] Beispielsweise ist das oben erwähnte manipulierte Video mit Nancy Pelosi hiervon nicht erfasst.[65]

Die beträchtlich voneinander abweichenden Standards der großen globalen Social-Media-Plattformen schaffen ein unübersichtliches Terrain, das organisierte Kampagnen leicht zur Distribution von Deepfakes ausnutzen können. Daher ist es zwar positiv, dass die Plattformen anscheinend die Notwendigkeit entsprechender Maßnahmen eingesehen haben. Aufgrund ihrer Uneinheitlichkeit werden diese Abwehrmechanismen gegen den Missbrauch von Deepfakes aber wohl wirkungslos bleiben. Dies wiederum unterstreicht, dass demokratische Regierungen eine Reihe von Mindeststandards für die Verteidigungsmechanismen und -systeme festlegen müssen, die Social-Media-Plattformen einsetzen sollten, um die Ausnutzung von Lücken in den jeweiligen Bestimmungen zu vermeiden.[66]

Deutschland hat mit dem Netzwerkdurchsetzungsgesetz (NetzDG)[67] bereits ein Rechtsinstrument geschaffen, mit dem gegen die Verbreitung von Deepfakes – einschließlich zum Zweck der Einflussnahme auf politische Prozesse verbreiteten Materials –vorgegangen werden kann und das Mindeststandards festlegt. Der dem Gesetz zugrunde liegende *Notice-and-Takedown*-Mechanismus entfaltet jedoch nur begrenzte Wirkung, da er die Plattformanbieter/-innen nicht zu proaktiven Maßnahmen verpflichtet, wie eine Anfang 2020 durchgeführte Studie des CEP zeigt.[68] Dennoch bietet der gegenwärtige Prozess zur Änderung des Gesetzes die Option, strengere Vorschriften einzuführen, um der Bedrohung durch Deepfakes zu begegnen.[69] Beispielsweise könnten Deepfake-Videos, die zu kriminellen oder politischen Zwecken missbraucht werden, als Verletzung des Rechts des Opfers am eigenen Bild definiert werden, sodass sie illegale Inhalte gemäß § 1 Abs. 3 NetzDG darstellen. Auf diese Weise wären sie bereits durch das Gesetz abgedeckt.[70]

Der derzeitige *Notice-and-Takedown*-Mechanismus basiert jedoch auf der Prämisse, dass Nutzer/-innen bösartige Inhalte identifizieren und Plattformen benachrichtigen, die dann verpflichtet sind, solche Inhalte zu entfernen, wenn festgestellt wird, dass sie gegen das NetzDG verstoßen.[71] Dieses Prinzip erscheint zu schwach, um die Verbreitung von Deepfake-Material, auch solchem, das den politischen Diskurs zu beeinflussen bezweckt, zu verhindern. Dies gilt insbesondere für Fälle, die eine rasche Reaktion erfordern, wie z. B. im Vorfeld einer Wahl oder einer entscheidenden Abstimmung lancierte Beeinflussungsversuche. Es muss auch darauf hingewiesen werden, dass die meisten Social-Media-Beiträge in den ersten Stunden nach ihrer Veröffentlichung die größte Beachtung finden. Ein System, das sich darauf verlässt, dass die Nutzer manipulative Inhalte kennzeichnen und die jeweilige Plattform benachrichtigen, die dann die Beschwerden prüft und erst danach unzulässige Inhalte entfernt, wird daher höchstwahrscheinlich nicht in der Lage sein, eine Wirkung durch Deepfakes zu verhindern.

Ohne spezielle Software sind hochwertige Deepfake-Videos nur schwierig, wenn nicht gar unmöglich, von authentischen Videos zu unterscheiden. Solche Tools sind für den/die Durchschnittsnutzer/-in im Allgemeinen nicht verfügbar. Plattformen, insbesondere solche mit globaler Reichweite, stehen daher in der Verantwortung, proaktive Maßnahmen zu ergreifen. Diese Unternehmen sind besonders gefährdet, da Deepfakes, die auf ihren Plattformen veröffentlicht werden, ihren großen Nutzerkreis wahrscheinlich einer wirksamen politischen Manipulation aussetzen werden. Daher sollten diese Plattformen verpflichtet werden, ihre Abwehrsysteme proaktiv zu stärken.

Zudem sollten weitere technikgestützte Maßnahmen ergriffen werden, um die Gesamtauswirkungen versuchter Einflussnahme auf politische Prozesse durch Deepfakes in Deutschland zu begrenzen.

5.2 Technik

Das Zersetzende an Deepfake-Technik, die als Mittel der Beeinflussung politischer Prozesse eingesetzt wird, liegt in ihrem Potenzial, die öffentliche Wahrnehmung der Realität und damit das öffentliche Vertrauen zu untergraben.[72] Für eine Abwehr gegen solche Angriffe sind Lösungen, die die Schaffung eines Fundus bestätigter Originalaufnahmen unterstützen, sowie eine Technik zur forensischen Überprüfung von Deepfake-Videos in Betracht zu ziehen.

5.2.1 Zertifizierung von Originalinhalten

Im Allgemeinen werden Videos als Wiedergabe der Realität wahrgenommen: „Das bewegte Bild wurde zum Versprechen einer reineren Wahrheit, einer Wahrheit, die nicht durch eine mögliche Veränderung, wie sie bei Standbildern möglich ist, verfälscht werden konnte.“[73] Deepfakes untergraben diese tief verwurzelte Überzeugung systematisch. Folglich könnten potenzielle technische Lösungen mit dem Ziel, einen Fundus verifizierter Originalaufnahmen zu schaffen, ein wichtiger Schritt sein. Ein solches Unterfangen würde natürlich einen langfristigen industriellen Wandel bei der zum Betrieb von Aufzeichnungsgeräten verwendeten Technik erfordern.

Wie oben erwähnt kann die *Control-Capture*-Technik ein wirksamer Ansatz sein.[74] Wenn bei der ursprünglichen Aufnahme des Materials ein digitales Wasserzeichen in Form eines *Hashs*[75] eingefügt werden kann, würde jede Manipulation des Materials automatisch den begleitenden *Hash* aktualisieren und anzeigen, dass das Video verändert wurde. Die *Hashing*-Technik ist ein bewährtes Prüfverfahren, das regelmäßig eingesetzt wird, um die Echtheit von Datensätzen nach einer Übertragung zu bestätigen.[76] Auch die *Blockchain*-Technik[77] kann eine unterstützende Rolle als Speicherort für „Original“-*Hashs* spielen, insbesondere für Aufnahmen von besonderer Bedeutung für die Allgemeinheit, wie z. B. Erklärungen von Regierungschefs zu wichtigen politischen Fragen usw.[78] Ein solcher offener Fundus mit *Hashs*, die mit Originalaufnahmen verknüpft sind, könnte dann dazu verwendet werden, Änderungen an den jeweiligen Aufnahmen rasch und zuverlässig zu überprüfen.

5.2.2 Unterstützung bei der Entwicklung von Techniken zur Deepfake-Erkennung

Die Entwicklung zuverlässiger Techniken zur Deepfake-Erkennung ist ein zweites wichtiges technikbasiertes Element. Solche Techniken werden für die forensische Analyse möglicher Deepfake-Videos von entscheidender Bedeutung sein. Darüber hinaus würde diese Entwicklung das notwendige Fachwissen schaffen, das es den Regulierungsbehörden ermöglicht, die Wirksamkeit der Abwehrsysteme einer Social-Media-Plattform für die Bekämpfung versuchter Einflussnahme auf politische Prozesse durch Deepfakes zu prüfen und zu bewerten.

In den USA gibt es mehrere aktuelle Initiativen zur Entwicklung von Techniken zur Deepfake-Erkennung. Mit der NDAA 2020 wurde ein nationaler Wettbewerb eingerichtet[79] und mehrere Unternehmen, darunter Facebook,[80] haben eigene Programme angekündigt. Microsoft, Facebook, Amazon Web Services (AWS) und die Partnership on AI[81] haben zudem 2019 die *Deep Fake Detection Challenge* ins Leben gerufen.[82]

Diese Initiativen zeigen, dass die Entwicklung einer solchen Technik nicht nur außerordentliches technisches Fachwissen, sondern auch erhebliche finanzielle Mittel erfordert. Deutschland befindet sich in einer guten Position, um zu ihrer Entwicklung beizutragen, da mehrere deutsche Forschungsinstitute bereits an KI und maschinellem Lernen arbeiten.[83]

5.3 Öffentliche Aufklärung

Es bedarf sowohl juristischer als auch technikbasierter Lösungen, um gegen die Einflussnahme auf politische Prozesse mittels Deepfakes vorzugehen. Diese werden natürlich nicht der Herausforderung gerecht, die sich aus der Tatsache ergibt, dass solche Beeinflussungsversuche möglicherweise bereits Wirkung zeigen und die öffentliche Wahrnehmung verändern. Die Risiken sind vor Wahlen oder parlamentarischen Abstimmungen besonders hoch. Leider werden daher forensische Ansätze, die darauf abzielen, Deepfakes zu entlarven, letztlich allein nicht wirksam sein.

2019 ergab eine Studie über die Online-Communities in mehreren europäischen Ländern im Vorfeld der Wahlen zum Europaparlament, dass

> *„die Reichweite von Faktenprüfern begrenzt ist, oft auf jene digitalen Gemeinschaften, die nicht Ziel von Desinformation sind oder diese verbreiten."*[84]

Dies lässt sich auch darauf zurückführen, dass sich Falschmeldungen in sozialen Medien schneller verbreiten als wahrheitsgemäße Informationen. Eine 2018 durchgeführte Längsschnitt-Forschungsstudie ergab:

> *Falsche Meldungen verbreiteten sich in allen Informationskategorien wesentlich weiter, schneller, tiefer und breiter als wahre, und die Auswirkungen waren bei politischen Falschmeldungen ausgeprägter als bei Falschmeldungen aus den Bereichen Terrorismus, Naturkatastrophen, Wissenschaft, urbanen Legenden oder Finanzinformationen.*[85]

Die Forscher/-innen kamen auch zu dem Schluss, dass menschliches Verhalten eine wichtigere Größe bei der Verbreitung von Falschmeldungen ist als automatisierte Systeme. Demnach würden Eingriffe in das Nutzungsverhalten, wie das Kennzeichnen von Falschmeldungen und eine Abschreckung vor deren Verbreitung, für das Eindämmen der Verbreitung von Falschmeldungen stärker wirken als andere Maßnahmen, wie etwa nachträgliche Eingriffe zur Faktenüberprüfung oder die Implementierung technischer Lösungen.[86]

Diese Erkenntnisse lassen sich auch auf Deepfakes anwenden, die – wenn sie zur Einflussnahme auf politische Prozesse instrumentalisiert werden – eine spezifische Unterkategorie von Falschmeldungen bilden. Ein unverzichtbares Abwehrinstrument ist daher die Aufklärung der Öffentlichkeit – eine gestärkte Internetkompetenz der Nutzerinnen und Nutzer.

Natürlich könnten sich einige Lösungen darauf konzentrieren, die sogenannte Gegenrede zu fördern, ein Konzept, das davon ausgeht, dass sich die Wahrheit in einem offenen Wettbewerb der Ideen schließlich

durchsetzen wird. Die Annahme, dass eine ungehemmte Verbreitung von Ideen Deepfakes zurückdrängen wird, beruht jedoch auf einem mangelnden Verständnis der technischen Mechanismen zur Verbreitung von Nachrichten in sozialen Medien. Die Algorithmen von Social-Media-Plattformen sind darauf ausgelegt, die auf der Plattform verbrachte Zeit zu verlängern. Dabei fördern sie unweigerlich auch die Neigung von Einzelpersonen zum Austausch falscher Informationen einschließlich Deepfakes. Daher haben sie auf diesem Marktplatz der Information eine zentrale Weichenfunktion, die die freie Verbreitung von Informationen mitunter negativ beeinflusst.

> *Bedingungen wie die strukturellen und wirtschaftlichen Veränderungen, die sich auf die Nachrichtenmedien ausgewirkt haben, die zunehmende Fragmentierung und Personalisierung sowie die zunehmend algorithmisch diktierte Verbreitung und Nutzung von Inhalten wirken sich daher auf die Produktion und den Fluss von Nachrichten in einer Weise aus, die die oben beschriebene Annahme, dass sich legitime Nachrichten systematisch gegen falsche Nachrichten durchsetzen, unterwandert.*[87]

Obwohl sie bedeutsam ist, sollte die Gegenrede von breiten öffentlichen Aufklärungskampagnen begleitet werden, die ein Bewusstsein dafür schaffen, dass Sehen keinen automatischen Glauben an Echtheit mehr begründet.[88] Dies könnte durch Bemühungen erreicht werden, die allgemeine Internetkompetenz zu erhöhen, beginnend mit spezialisierten Kursen als Teil des regulären Schullehrplans. In Deutschland gibt es in dieser Hinsicht bereits einige Bemühungen, die intensiviert werden sollten[89] und auch von Medienverbänden geleitet werden könnten. 2019 startete beispielsweise die kanadische NGO für Medienentwicklung Journalists for Human Rights (JHR) ihr von der kanadischen Regierung unterstütztes Projekt „Fighting Desinformation through Strengthened Media and Citizen Preparedness in Canada“ („Bekämpfung von Desinformation durch verstärkte Aufmerksamkeit von Medien und Bürgern in Kanada“).[90]

Ein Nebeneffekt der wachsenden öffentlichen Skepsis gegenüber Online-Inhalten ist jedoch, dass wahrscheinlich die sogenannte *liar's dividend* häufiger und wirksamer genutzt werden wird.[91]

Eine zweite notwendige Strategie zur Aufklärung der Öffentlichkeit sollte Bemühungen umfassen, den Bürger/-innen dabei zu helfen, zwischen den Mechanismen der Verbreitung von Inhalten, wie z. B. Social-Media-Plattformen, und den Mechanismen der Nachrichtenproduktion zu unterscheiden. Professionelle Medienorganisationen lassen sich in der Regel von Verhaltenskodizes leiten. Der Verhaltenskodex des Deutschen Presserats zum Beispiel umreißt in Abschnitt 1:

> *„Die Achtung vor der Wahrheit, die Wahrung der Menschenwürde und die wahrhaftige Unterrichtung der Öffentlichkeit sind oberste Gebote der Presse.“*[92]

Neben der Förderung des Bewusstseins und der Internetkompetenz könnten bestehende Funktionen, wie z. B. die umgekehrte Bildsuche, der Öffentlichkeit zur Verfügung gestellt und der Zugang zu diesen Funktionen für Social-Media-Plattformen verbindlich vorgeschrieben werden.

> *Die umgekehrte Bildsuche hat es Journalist/-innen, Faktenprüfer/-innen und gewöhnlichen Internetnutzer/-innen ermöglicht, Originalfotos ausfindig zu machen, aus denen Fälschungen hergestellt werden. Benutzer/ -innen können ein Bild hochladen und dann mithilfe von Computer-Vision ähnliche Fotos online entdecken, die das Foto als verändert oder als aus dem Zusammenhang gerissen kennzeichnen.*[93]

Die Förderung der Internetkompetenz und des öffentlichen Bewusstseins sollte in Verbindung mit einer verstärkten Verfügbarkeit von Instrumenten, wie z. B. der umgekehrten Bildsuche, Teil einer allgemeinen Aufklärungspolitik werden, die zur Stärkung der gesellschaftlichen Abwehrkräfte gegenüber der Beeinflussung politischer Prozesse durch Deepfakes eingesetzt werden könnte.

42 1994 nutzte der Film *Forest Gump* Deepfake-Videomanipulation besonders innovativ, indem die Hauptfigur, verkörpert durch den Schauspieler Tom Hanks, in mehrere historische Originalaufnahmen eingefügt wurde.

43 Zum Beispiel verwendete eine indische Partei Deepfake-Technik, um ihre Wahlkampfbotschaft in mehreren Sprachen zu verbreiten. *Siehe*: J. Fergus, Deepfake video in multiple languages is the first of its kind in an Indian election. It's a 'positive campaign', INPUT, 19. Februar 2020, https://www.inputmag.com/culture/a-deepfake-video-is-the-first-of-its-kind-in-indian-election-campaign [06.05.2020]. Systematische Ansätze, die den potenziellen Einsatz und Missbrauch dieser Technik umreißen, finden sich bei R. Chesney/D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security; E. Meskys/A. Liaudanskas/J. Kalpokiene/P. Jurcys, Regulating deep fakes: legal and ethical considerations. In: Journal of Intellectual Property Law & Practice 15/2020, Heft 1, S. 24–31, https://academic.oup.com/jiplp/article/15/1/24/5709090 [06.05.2020].

44 AB 730, Elections: deceptive audio or visual media, October 2019, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 [06.05.2020].

45 Ebd. Deutsche Übersetzung: *„Dieser Abschnitt gilt nicht für eine Internet-Website oder eine regelmäßig erscheinende Zeitung, Zeitschrift oder ein anderes Periodikum von allgemeiner Verbreitung, einschließlich einer Internet- oder elektronischen Publikation, die routinemäßig Nachrichten und Kommentare von allgemeinem Interesse enthält und die materiell täuschende Audio- oder visuelle Medien veröffentlicht, die durch diesen Abschnitt verboten sind, wenn in der Publikation klar angegeben ist, dass die materiell täuschenden Audio- oder visuellen Medien die Rede oder das Verhalten des Kandidaten nicht korrekt wiedergeben."*

46 A. Metwally/J. P. Mohler, Manipulated Media: Examining California's Deepfake Bill, JOLT Digest, 12. November 2019, http://jolt.law.harvard.edu/digest/manipulated-media-examining-californias-deepfake-bill [06.05.2020].

47 B. M. Nonnecke, Opinion: California's Anti-Deepfake Law Is Far Too Feeble. While well intentioned, the law has too many loopholes for malicious actors and puts too little responsibility on platforms, Wired, 5. November 2019, https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/ [06.05.2020].

48 D. Castro, State Government Might Not Be Enough to Stop Deepfakes, Governing, 7. Januar 2020, https://www.governing.com/news/headlines/State-Government-Might-Not-Be-Enough-to-Stop-Deepfakes.html [06.05.2020].

49 H. R. 3230 Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/3230/text [06.05.2020].

50 National Defense Authorization Act for Fiscal Year 2020, S. 1790, 116th Congress, 1st Session (2019), https://www.govinfo.gov/content/pkg/BILLS-116s1790enr/pdf/BILLS-116s1790enr.pdf [06.05.2020].

51 Ebd. Deutsche Übersetzung: *„(A) die potentiellen Auswirkungen maschinenmanipulierter Medien (allgemein als „Deepfakes" bekannt) auf die nationale Sicherheit; und (B) die tatsächliche oder potenzielle Nutzung maschinell manipulierter Medien durch ausländische Regierungen zur Verbreitung von Desinformation oder für andere böswillige Aktivitäten."*

52 Ebd. Deutsche Übersetzung: *„Eine aktualisierte Auflistung der Abwehrtechniken, die von der Regierung der Vereinigten Staaten oder vom privaten Sektor mit Unterstützung der Regierung entwickelt wurden oder entwickelt und eingesetzt werden könnten, um die Verwendung maschinell manipulierter Medien und maschinell erstellter Texte durch ausländische Regierungen, mit ausländischen Regierungen verbundene Unternehmen oder ausländische Einzelpersonen zu erschweren, aufzudecken und zuzuordnen, zusammen mit einer Analyse der Vorteile, Grenzen und Nachteile dieser benannten Abwehrtechniken einschließlich neuer potentieller Bedenken im Hinblick auf den Datenschutz."*

53 Ebd. Deutsche Übersetzung: *„um Erforschung, Entwicklung oder Kommerzialisierung von Techniken zur automatischen Erkennung maschinell manipulierter Medien zu fördern."*

54 Zum Beispiel gestattet die „Fair Use"-Doktrin eine Verwendung von Teilen urheberrechtlich geschützten Materials für öffentliche Zwecke wie Kommentare, Satire, Parodien, Berichte, Bildung oder Forschung.

55 Abschnitt 230 (c) (1) und (2) des Telecommunications Act 1996, Pub. LA. No. 104–104, 110 Stat. 56 (1996), https://transition.fcc.gov/Reports/tcom1996.pdf [06.05.2020]. Deutsche Übersetzung: *„Kein Anbieter oder Benutzer eines interaktiven Computerdienstes darf als Herausgeber oder Mittler von Informationen behandelt werden, die durch einen anderen Anbieter von Informationsinhalten bereitgestellt werden. […] Kein Anbieter oder Benutzer eines interaktiven Computerdienstes kann haftbar gemacht werden für – (A) eine Handlung, die freiwillig und in gutem Glauben unternommen wird, um den Zugang zu oder die Verfügbarkeit von Material einzuschränken, das der Anbieter oder Nutzer als obszön, unzüchtig, lüstern, schmutzig, übermäßig gewalttätig, belästigend oder anderweitig anstößig erachtet, unabhängig davon, ob dieses Material verfassungsrechtlich geschützt ist oder nicht; oder (B) eine Maßnahme, die ergriffen wird, um Anbietern von Informationsinhalten oder anderen die technischen Mittel zur Beschränkung des Zugangs zu dem in Absatz (1) beschriebenen Material zu ermöglichen oder zur Verfügung zu stellen."*

56 M. Bickert, Enforcing Against Manipulated Media, Facebook, 6. Januar 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [06.05.2020].

57 Y. Roth/A. Achuthan, Building rules in public: Our approach to synthetic & manipulated media, Twitter, 4. Februar 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [06.05.2020].

58 L. Mahendran/N. Alsherif, Adding clarity to our Community Guidelines, TikTok, 8. Januar 2020, https://newsroom.tiktok.com/en-us/adding-clarity-to-our-community-guidelines [06.05.2020].

59 Updates to Our Policy Around Impersonation, Reddit, 9. Januar 2020, https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/ [06.05.2020].

60 How YouTube supports elections, YouTube, 3. Februar 2020, https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html?m=1 [06.05.2020].

61 Das Unternehmen nahm diese Richtlinienänderung vor, während es laut Berichten selbst an der Deepfake-Technik arbeitete (siehe J. Constine, ByteDance & TikTok have secretly built a deepfakes maker, TechCrunch, 3. Januar 2020, https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/ [06.05.2020]).

62 Community-Richtlinien, TikTok, Januar 2020, https://www.tiktok.com/community-guidelines?lang=en [06.05.2020].

63 Community Standards, Manipulated Media, Facebook, https://www.facebook.com/communitystandards/manipulated_media [06.05.2020].

64 Siehe zum Beispiel G. Edelman, Facebook's Deepfake Ban Is a Solution to a Distant Problem. The platform has a plan to deal with tomorrow's disinformation. But what about today's? Wired, 7. Januar 2020, https://www.wired.com/story/facebook-deepfake-ban-disinformation/ [06.05.2020]; J. Sachs, Facebook's Ban On Deepfakes Not Likely To Help Stop Spread Of Misinformation, Grit Daily, 8. Januar 2020, https://gritdaily.com/ban-on-deepfakes-facebook/ [06.05.2020]; A. Khalid, Facebook's deepfake ban ignores most visual misinformation, Quartz, 9. Januar 2020, https://qz.com/1781809/facebooks-deepfake-ban-wont-remove-most-visual-misinformation/ [06.05.2020].

65 Siehe Unterkapitel 4.2.

66 In den Vereinigten Staaten haben bereits mehrere Senatoren öffentlich gefordert, dass die Technologiebranche diesbezügliche Standards festlegt (siehe B. Vincent, Sens. Marco Rubio and Mark Warner want Facebook, YouTube, TikTok and others to create industry standards for handling synthetic content, Nextgov, 2. Oktober 2019, https://www.nextgov.com/emerging-tech/2019/10/lawmakers-press-social-media-giants-confront-deepfake-threats/160325/ [06.05.2020]).

67 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG), https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html [06.05.2020].

68 A. Ritzmann/M. Macori/H.-J. Schindler, NetzDG 2.0. Empfehlungen zur Weiterentwicklung des Netzwerkdurchsetzungsgesetzes (NetzDG) und Untersuchung zu den tatsächlichen Sperr- und Löschprozessen von YouTube, Facebook und Instagram, Counter Extremism Project, 12. März 2020, https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper.pdf [06.05.2020].

69 Bundesregierung, Entwurf eines Gesetzes zur Änderung des Netzwerkdurchsetzungsgesetzes, 31. März 2020, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RegE_Aenderung_NetzDG.pdf?__blob=publicationFile&v=2 [06.05.2020].

70 H.-J. Schindler/N. Semaan, Democratising Deepfakes. How Technological Development Can Influence Our Social Consensus. In: International Reports of the Konrad Adenauer Stiftung 1/2020, S. 60–68, https://www.kas.de/documents/259121/8620647/Democratising+Deepfakes.pdf/8b3a9ba0-b2ff-2e8d-32be-f7992894a5e5?version=1.0&t=1585317007608 [06.05.2020].

71 § 1 Abs. 3 NetzDG, https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html [06.05.2020].

72 W. A. Galston, Is seeing still believing? The deepfake challenge to truth in politics, Brookings Institution, 8. Januar 2020, https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/ [06.05.2020].

73 T. Le Wagner/A. Blewer, 'The Word Real Is No Longer Real': Deepfakes, Gender, and the Challenges of AI-Altered Video. In: Open Information Science 3/2019, S. 33, https://www.researchgate.net/publication/334730810_The_Word_Real_Is_No_Longer_Real_Deepfakes_Gender_and_the_Challenges_of_AI-Altered_Video [06.05.2020].

74 Siehe Unterkapitel 3.2.

75 Hashing ist definiert als die Generierung eines Werts oder von Werten aus einer Textfolge mittels einer mathematischen Funktion (siehe Definition Hashing, Techopedia, 21. November 2017, https://www.techopedia.com/definition/14316/hashing [06.05.2020]).

76 Definition Hash, Tech Terms, 21. April 2018, https://techterms.com/definition/hash [06.05.2020].

77 Eine Blockchain (Blockkette) oder verteiltes Ledger ist die offene dezentralisierte Verteilung kryptografisch gesicherter Hashs, die von einem Peer-to-Peer-Netzwerk verwaltet wird. Dies ermöglicht einen offenen Zugang für alle Teilnehmer, während die für die Hashs verwendete Verschlüsselungsmethode sicherstellt, dass diese nicht verändert werden können (siehe Blockchain. What Is Blockchain Technology? How Does Blockchain Work?, BuiltIn, https://builtin.com/blockchain [06.05.2020]).

78 A. G. Martinez, The Blockchain Solution to Our Deepfake Problems. Technology to hack videos will only keep getting better. A decentralized ledger might help us know when we're seeing the truth, Wired, 26. März 2018, https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/ [06.05.2020]; P. Madsen, Combating deepfakes with distributed ledgers, Hedera Hashgraph, 10. Juni 2019, https://www.hedera.com/blog/using-distributed-ledgers-to-combat-deepfakes [06.05.2020].

79 Siehe Abschnitt 5.2.1.

80 M. Schroepfer, Creating a data set and a challenge for deepfakes, Facebook AI, 5. September 2019, https://ai.facebook.com/blog/deepfake-detection-challenge/ [06.05.2020].

81 Partnership of AI ist eine Multi-Stakeholder-Organisation von Unternehmen und Experten für künstliche Intelligenz (siehe Frequently Asked Questions, Partnership on AI, https://www.partnershiponai.org/faq/ [06.05.2020]).

82 Deepfake Detection Challenge, https://deepfakedetectionchallenge.ai/ [06.05.2020].

83 Siehe zum Beispiel Nationales Forschungszentrum für angewandte Cybersicherheit ATHENE, https://www.athene-center.de/; Max-Planck-Institut für Intelligente Systeme, https://www.is.mpg.de/; Fraunhofer-Allianz Big Data und Künstliche Intelligenz, https://www.fraunhofer.de/en/institutes/institutes-and-research-establishments-in-germany/fraunhofer-alliances/big-data-and-artificial-intelligence-alliance.html; Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme, https://www.iais.fraunhofer.de/en/research/artificial-intelligence.html; Max-Planck-Institut für Maschinelles Lernen, https://www.cis.mpg.de/machine-learning/. Einen nützlichen Überblick über die laufenden Forschungsaktivitäten in Deutschland findet sich hier: Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Manuel Höferlin, Frank Sitta, Grigorios Aggelidis, weiterer Abgeordneter und der Fraktion der FDP – Drucksache 19/15210 – Beschäftigung der Bundesregierung mit Deepfakes, 2. Dezember 2019, http://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf [06.05.2020].

84 How Effective Are Fact-Checkers? A preliminary analysis on how successful fact-checkers are at disseminating content across different digital communities of opinion, Alto Analy-

tics, 12. Juli 2019, https://www.alto-analytics.com/en_US/fact-checkers/ [06.05.2020].

85 S. Vosoughi/D. Roy/S. Aral, The spread of true and false news online. In: Science 359/2018, S. 1146, https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf [06.05.2020]. Die Forscher analysierten rund 126.000 Geschichten, die zwischen 2006 und 2017 von rund 3 Millionen Menschen mehr als 4,5 Millionen Mal getwittert wurden.

86 S. Vosoughi/D. Roy/S. Aral, The spread of true and false news online, S. 1150.

87 P. M. Napoli, What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble. In: Federal Communications Law Journal 70/2017–2018, Heft 1, S. 59, http://www.fclj.org/wp-content/uploads/2018/04/70.1-Napoli.pdf [06.05.2020].

88 Siehe zum Beispiel I. Beridze/J. Butcher, When seeing is no longer believing. In: Nature Machine Intelligence 1/2019, S. 332–334, https://www.nature.com/articles/s42256-019-0085-5 [27.05.2020]; H. K. Hall, Deepfake Videos: When Seeing Isn't Believing. In: Catholic University Journal of Law and Technology 27/2018, Heft 1, S. 51–76 [06.05.2020].

89 Siehe zum Beispiel Digitally Educated, deutschland.de, 6. Februar 2018, https://www.deutschland.de/en/topic/knowledge/digital-literacy-for-school-pupils-three-good-examples [06.05.2020].

90 Journalists for Human Rights (JHR), Launching JHR's program on 'Fighting Disinformation through Strengthened Media and Citizen Preparedness in Canada', CISION, 27. September 2019, https://www.newswire.ca/news-releases/launching-jhr-s-program-on-fighting-disinformation-through-strengthened-media-and-citizen-preparedness-in-canada--899686785.html [06.05.2020].

91 P. Chadwick, The liar's dividend, and other challenges of deep-fake news, Guardian, 22. Juli 2018, https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin [06.05.2020].

92 Publizistische Grundsätze (Pressekodex). Richtlinien für die publizistische Arbeit nach den Empfehlungen des Deutschen Presserats, S. 2, https://www.presserat.de/files/presserat/dokumente/download/Pressekodex2017light_web.pdf [12.06.05.2020]. Der Deutsche Presserat ist auch das zuständige Gremium für die Behandlung von Beschwerden gegen Journalisten und Medienorganisationen, die gegen den Pressekodex verstoßen (siehe ebd., S. 12–15).

93 A. Engler, Fighting deepfakes when detection fails, Brookings Institution, 14. November 2019, https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/ [06.05.2020].

6. Fazit

Dieser Bericht skizziert den aktuellen Sachstand bezüglich Deepfake-Videos und veranschaulicht, wie die zunehmende Verbreitung dieser Technik den gesellschaftlichen Zusammenhalt untergraben kann. Der rasche technische Fortschritt eröffnet neue Chancen bei der Produktion von Deepfakes und bietet böswilligen Akteur/-innen – einschließlich staatlichen – neue Möglichkeiten für schädliche Eingriffe in den politisch-gesellschaftlichen Diskurs. Gegenwärtig scheint die hohe Menge an Trainingsdaten, die für die Herstellung hochentwickelter Deepfakes erforderlich ist, die Massenproduktion einzuschränken. Jüngste Fortschritte, die die Produktion aus einem einzigen Standbild synthetisierter Bewegtbilder ermöglichen, deuten jedoch darauf hin, dass auch diese Hürde in naher Zukunft wegfallen könnte.[94]

Gegenwärtig ist die Verwendung von Deepfakes in einem kriminellen Kontext nach wie vor weit verbreitet, insbesondere bei nicht einvernehmlicher Pornografie. Bekannte Fälle von Deepfakes, die zum Zwecke der Beeinflussung politischer Prozesse hergestellt wurden, betrafen Videos mit relativ leicht erkennbaren Manipulationen. Dennoch verursachte selbst dieses minderwertige Material politische Turbulenzen. Sollte diese Technik im Rahmen einer von böswilligen staatlichen Akteur/-innen gesteuerten, organisierten und zielgerichteten Kampagne zur Desinformation massenhaft eingesetzt werden, könnten die zersetzenden Auswirkungen einer solchen Kampagne erheblich sein.

Die in der Deepfake-Technik begründete langfristige Herausforderung liegt in der allmählichen Erosion des öffentlichen Vertrauens, da sie das etablierte Wahrheitsverständnis der Gesellschaft untergräbt. Letztlich stellt der Missbrauch dieser Technik eine Bedrohung für die freie und offene politische Debatte dar. Das diskursive Prinzip, das der deutsche Philosoph Jürgen Habermas als eines der Kernelemente einer Demokratie definiert,[95] wird untergraben, wenn der Diskurs selbst auf manipulierten Wahrnehmungen der Wirklichkeit beruht.

In diesem Sinne sind Deepfakes der technisch fortschrittlichste Aspekt der ständig zunehmenden Bedrohung durch gefälschte Nachrichten (*fake news*). Die zunehmende Verfügbarkeit kostengünstiger globaler Verbreitungsmechanismen – vor allem Social-Media-Plattformen und ihre zunehmend zentrale Rolle bei der Bereitstellung von Informationen für die Öffentlichkeit – verschärft diese Gefahr noch. Der unregulierte Charakter dieser Plattformen und ihre stark divergierenden Unternehmenspolitiken sind bedeutende Herausforderungen bei der Verteidigung des politischen Diskurses und der Aufrechterhaltung des sozialen Zusammenhalts.

Folglich müssen wir dringend eine strategische Diskussion beginnen, die auf die Entwicklung eines wirksamen Abwehransatzes gegen diese aufkommende Bedrohung abzielt. Derzeit ist Deutschland von dieser Technik nicht so stark betroffen wie andere Länder. Das deutet darauf hin, dass sich Deutschland bei der Bewältigung dieser neuen Herausforderung in einem frühen Stadium befindet und noch genügend Zeit bleibt, wirksame Antworten zu entwickeln. Es ist unwahrscheinlich, dass einzelne isolierte Maßnahmen Wirkung zeigen. Die durch Deepfakes fortgeschrittene, intensivierte Einflussnahme auf politische Prozesse stellt eine komplexe Herausforderung dar, deren Bewältigung eines mehrgleisigen Ansatzes bedarf, der sich gegenseitig ergänzende juristische Bestimmungen, technische Lösungen und Maßnahmen zur Aufklärung der Öffentlichkeit kombiniert.

Die in diesem Bericht umrissenen Elemente und Maßnahmen erfordern eine politische Debatte, die jetzt beginnen muss. Die Konrad-Adenauer-Stiftung und das Counter Extremism Project hoffen, mit diesem Bericht einen Beitrag zur Eröffnung dieser entscheidenden Debatte zu leisten.

94 M. Weisberger, This Animated Mona Lisa Was Created by AI, and It Is Terrifying, Live Science, 27. Mai 2019, https://www.livescience.com/65573-mona-lisa-deepfakes.html [06.05.2020].

95 J. Habermas, The Theory of Communicative Action. Vol. I: Reason and the Rationalization of Society, T. McCarthy (trans.). Boston: Beacon, 1984.

7. Literaturverzeichnis

A **Ajder, Henry / Patrini, Giorgio / Cavalli, Francesco / Cullen, Laurence,** *The State of Deep Fakes. Landscape, Threats and Impact*, Deeptrace, September 2019, https://regmedia.co.uk/2019/10/08/deepfake_report.pdf [06.05.2020].

Alto Analytics, *How Effective Are Fact-Checkers? A preliminary analysis on how successful fact-checkers are at disseminating content across different digital communities of opinion*, 12. Juli 2019, https://www.alto-analytics.com/en_US/fact-checkers/ [06.05.2020].

ARD/ZDF *Online-Studie 2019*, http://www.ard-zdf-onlinestudie.de/files/2019/ARD-ZDF-Onlinestudie-Grafik-2019.pdf [06.05.2020].

B **Bakowski, Piotr / Puccio, Laura,** *Foreign fighters – Member State responses and EU action*, Wissenschaftlicher Dienst des Europäischen Parlaments, März 2016, https://www.europarl.europa.eu/EPRS/EPRS-Briefing-579080-Foreign-fighters-rev-FINAL.pdf [06.05.2020].

Benková, Lívia, *The Rise of Russian Disinformation in Europe*. In: Austria Institut für Europa- und Sicherheitspolitik, Fokus 3/2018, https://www.aies.at/download/2018/AIES-Fokus_2018-03.pdf [06.05.2020].

Beridze, Irakli / Butcher, James, *When seeing is no longer believing*. In: Nature Machine Intelligence 1/2019, S. 332–334, https://www.nature.com/articles/s42256-019-0085-5 [27.05.2020].

Bickert, Monika, *Enforcing Against Manipulated Media*, Facebook, 6. Januar 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [06.05.2020].

Bradshaw, Samantha / Howard, Philip N., *The Global Disinformation Disorder*: 2019 Global Inventory of Organised Social Media Manipulation. Working Paper 2/2019, Oxford, UK: Project on Computational Propaganda, https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf [06.05.2020].

Breland, Ali, *The Bizarre and Terrifying Case of the „Deepfake" Video that Helped Bring an African Nation to the Brink*, MotherJones, 15. März 2019, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [06.05.2020].

BuiltIn, Blockchain. *What Is Blockchain Technology? How Does Blockchain Work?*, https://builtin.com/blockchain [06.05.2020].

Bundesregierung, *Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Manuel Höferlin, Frank Sitta, Grigorios Aggelidis, weiterer Abgeordneter und der Fraktion der FDP* – Drucksache 19/15210 – Beschäftigung der Bundesregierung mit Deepfakes, 2. Dezember 2019, http://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf [06.05.2020].

Burchard, Hans von der, *Belgian socialist party circulates 'deep fake' Donald Trump video*, Politico, 21. Mai 2018, https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/ [06.05.2020].

C **Cahalan, Sarah,** *How misinformation helped spark an attempted coup in Gabon*, The Washington Post, 13. Februar 2020, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/ [06.05.2020].

Castro, Daniel, *State Government Might Not Be Enough to Stop Deepfakes*, Governing, 7. Januar 2020, https://www.governing.com/news/headlines/State-Government-Might-Not-Be-Enough-to-Stop-Deepfakes.html [06.05.2020].

Chadwick, Paul, *The liar's dividend, and other challenges of deep-fake news*, Guardian, 22. Juli 2018, https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin [06.05.2020].

Chesney, Robert / Citron, Danielle Keats, *Deep fakes: A looming challenge for privacy, democracy, and national security*. In: California Law Review 107/2019, S. 1753–1819, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 [06.05.2020].

Coats, Daniel R., *Statement for the Record: Worldwide Threat Assessment of the Intelligence Community*, 29. Januar 2019, https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf [06.05.2020].

Constine, Josh, *ByteDance & TikTok have secretly built a deepfakes maker*, Techcrunch, 3. Januar 2020, https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/ [06.05.2020].

Coyne, Bridget, *Helping identify 2020 US election candidates on Twitter*, Twitter, 12. Dezember 2019, https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html [06.05.2020].

D ***Digitally Educated,*** deutschland.de, 6. Februar 2018, https://www.deutschland.de/en/topic/knowledge/digital-literacy-for-school-pupils-three-good-examples [06.05.2020].

E **Edelman, Gilad,** *Facebook's Deepfake Ban Is a Solution to a Distant Problem. The platform has a plan to deal with tomorrow's disinformation. But what about today's?*, Wired, 7. Januar 2020, https://www.wired.com/story/facebook-deepfake-ban-disinformation/ [06.05.2020].

Engler, Alex, *Fighting deepfakes when detection fails*, Brookings, 14. November 2019, https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/ [06.05.2020].

F **Farid, Hany,** *Photo Forensics*. Cambridge, MA/London: MIT Press. 2016.

Fergus, J., *Deepfake video in multiple languages is the first of its kind in an Indian election. It's a 'positive campaign'*, INPUT, 19. Februar 2020, https://www.inputmag.com/culture/a-deepfake-video-is-the-first-of-its-kind-in-indian-election-campaign [06.05.2020].

Fichera, Angelo, *Manipulated Video Targeting Pelosi Goes Viral*, FactCheck.Org, 24. Mai 2019, https://www.factcheck.org/2019/05/manipulated-video-targeting-pelosi-goes-viral/ [06.05.2020].

G **Galston, William A.,** *Is seeing still believing? The deepfake challenge to truth in politics*, Brookings, 8. Januar 2020, https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/ [06.05.2020].

H **Habermas, Jürgen,** *The Theory of Communicative Action*. *Vol. I: Reason and the Rationalization of Society,* T. McCarthy (trans.). Boston: Beacon, 1984.

Hale, James, *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*, Tubefilter, 7. Mai 2019, https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/ [06.05.2020].

Hall, Holly Kathleen, Deepfake Videos: *When Seeing Isn't Believing*. In: Catholic University Journal of Law and Technology 27/2018, Heft 1, S. 51–76, https://scholarship.law.edu/cgi/viewcontent.cgi?article=1060&context=jlt [06.05.2020].

Harwell, Drew, *Faked Pelosi videos, slowed to make her appear drunk, spread across social media,* The Washington Post, 24. März 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [06.05.2020].

Hölig, Sascha / Hasebrink, Uwe, *Nachrichtennutzung über soziale Medien im internationalen Vergleich.* In: Media Perspektiven 11/2016, S. 534–548, https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2016/11-2016_Hoelig_Hasebrink.pdf [06.05.2020].

J **Journalists for Human Rights (JHR),** Vorstellung des JHR-Programms *„Fighting Disinformation through Strengthened Media and Citizen Preparedness in Canada"*, CISION, 27. September 2020, https://www.newswire.ca/news-releases/launching-jhr-s-program-on-fighting-disinformation-through-strengthened-media-and-citizen-preparedness-in-canada--899686785.html [06.05.2020].

K **Karlsefni, Thorfinn,** *Nicholas Cage, Sound of Music (Deepfake)*, https://youtu.be/MHkZEpfUnAA [06.05.2020].

Karras, Tero / Laine, Samuli / Aila, Timo, *A style-based generator architecture for generative adversarial networks.* In: IEEE Conference on Computer Vision and Pattern Recognition, 2019, S. 4401–4410.

Karras, Tero / Laine, Samuli / Aittala, Miika / Hellsten, Janne / Lehtinen, Jaakko / Aila, Timo, *Analyzing and improving the image quality of stylegan,* 2019, https://arxiv.org/abs/1912.04958 [06.05.2020].

Khalid, Amrita, *Facebook's deepfake ban ignores most visual misinformation*, Quartz, 9. Januar 2020, https://qz.com/1781809/facebooks-deepfake-ban-wont-remove-most-visual-misinformation/ [06.05.2020].

L **Li, Yuezun / Chang, Ming-Ching / Lyu, Siwei,** *In ictu oculi: Exposing AI created fake videos by detecting eye blinking.* In: IEEE International Workshop on Information Forensics and Security, 2018, S. 1–7, https://arxiv.org/pdf/1806.02877.pdf [06.05.2020].

Lossau, Norbert, *Deep Fake: Gefahren, Herausforderungen und Lösungswege,* Konrad-Adenauer-Stiftung, Analysen & Argumente Nr. 382/2020, https://www.kas.de/documents/252038/7995358/AA382+Deep+Fake.pdf/de479a86-ee42-2a9a-e038-e18c208b93ac?version=1.0&t=1581576967612 [06.05.2020].

Lytvynenko, Jane, *A Belgian Political Party is Circulating a Trump Deepfake Video*, BuzzFeed, 20. Mai 2018, https://www.buzzfeednews.com/article/janelytvynenko/a-belgian-political-party-just-published-a-deepfake-video [06.05.2020].

M **Mahendran, Lavanya / Alsherif, Nasser,** *Adding clarity to our Community Guidelines*, TikTok, 8. Januar 2020, https://newsroom.tiktok.com/en-us/adding-clarity-to-our-community-guidelines [06.05.2020].

Madsen, Paul, *Combating deepfakes with distributed ledgers*, Hedera Hashgraph, 10. Juni 2019, https://www.hedera.com/blog/using-distributed-ledgers-to-combat-deepfakes [06.05.2020].

Martínez, Antonio García, *The Blockchain Solution to Our Deepfake Problems. Technology to hack videos will only keep getting better. A decentralized ledger might help us know when we're seeing the truth,* Wired, 26. März 2018, https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/ [06.05.2020].

Meskys, Edvinas / Liaudanskas, Aidas / Kalpokiene, Julija / Jurcys, Paulius, *Regulating deep fakes: legal and ethical considerations.* In: Journal of Intellectual Property Law & Practice, 15/2020, Heft 1, S. 24–31, https://academic.oup.com/jiplp/article/15/1/24/5709090 [06.05.2020].

Metwally, Amre / Mohler, J. P., *Manipulated Media: Examining California's Deepfake Bill,* JOLT Digest, 12. November 2019, http://jolt.law.harvard.edu/digest/manipulated-media-examining-californias-deepfake-bill [06.05.2020].

Metz, Rachel, *These people do not exist. Why websites are churning out fake images of people (and cats)*, CNN, 28. Februar 2020, https://edition.cnn.com/2019/02/28/tech/ai-fake-faces/index.html [06.05.2020].

N **Napoli, Philip M.,** *What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble.* In: Federal Communications Law Journal 70/2017–2018, Heft 1, S. 55–104, http://www.fclj.org/wp-content/uploads/2018/04/70.1-Napoli.pdf [06.05.2020].

Nonnecke, Brandie M., *Opinion: California's Anti-Deepfake Law Is Far Too Feeble. While well intentioned, the law has too many loopholes for malicious actors and puts too little responsibility on platforms.* Wired, 5. November 2019, https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/ [06.05.2020].

Noyes, Dan, *The Top 20 Valuable Facebook Statistics – Updated January 2020*, Zephoria, https://zephoria.com/top-15-valuable-facebook-statistics/ [06.05.2020].

O **Oord, Aaron van den / Dieleman, Sander / Zen, Heiga / Simonyan, Karen / Vinyals, Oriol / Graves, Alex / Kalchbrenner, Nal / Senior, Andrew / Kavukcuoglu, Koray,** *Wavenet: A generative model for raw audio,* 2016, https://arxiv.org/abs/1609.03499 [06.05.2020].

O'Sullivan, Donie, *A high school student created a fake 2020 candidate. Twitter verified it,* CNN, 28. Februar 2020, https://edition.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html [06.05.2020].

P **Paris, Britt / Donovan, Joan,** *Deepfakes and Cheap Fakes. The Manipulation of Audio and Visual Evidence*, 2019, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf [27.05.2020].

Pearson, Helen, *Image manipulation: CSI: Cell biology.* In: Nature 434/2005, S. 952–953, https://www.nature.com/articles/434952a.pdf?proof=true&draft=collection%3Fproof%3Dtrue [06.05.2020].

Presserat, *German Press Code*, https://www.presserat.de/files/presserat/dokumente/download/Press%20Code.pdf [06.05.2020].

R **Reddit,** Updates to Our Policy Around Impersonation, 9. Januar 2020, https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/ [06.05.2020].

Ritzmann, Alexander / Macori, Marco / Schindler, Hans-Jakob, *NetzDG 2.0. Empfehlungen zur Weiterentwicklung des Netzwerkdurchsetzungsgesetzes (NetzDG) und Untersuchung zu den tatsächlichen Sperr- und Löschprozessen von YouTube, Facebook und Instagram,* CEP-Strategiepapier, 12. März 2020, https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper.pdf [06.05.2020].

Roose, Kevin / Frenkel, Sheera / Perlroth, Nicole, *Facebook, Google and Twitter Struggle to Handle November's Election,* The New York Times, 29. März 2020, https://www.bizjournals.com/philadelphia/news/2020/03/30/facebook-google-and-twitter-struggle-to-handle.html [06.05.2020].

Roth, Yoel / Achuthan, Ashita, *Building rules in public: Our approach to synthetic & manipulated media,* Twitter, 4. Februar 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [06.05.2020].

S **Sachs, Julia,** *Facebook's Ban On Deepfakes Not Likely To Help Stop Spread Of Misinformation,* Grit Daily, 8. Januar 2020, https://gritdaily.com/ban-on-deepfakes-facebook/ [06.05.2020].

Schindler, Hans-Jakob / Semaan, Nauel, Democratising Deepfakes. *How Technological Development Can Influence Our Social Consensus*. In: International Reports of the Konrad-Adenauer-Stiftung 1/2020, S. 60–68, https://www.kas.de/documents/259121/8620647/Democratising+Deepfakes.pdf/8b3a9ba0-b2ff-2e8d-32be-f7992894a5e5?version=1.0&t=1585317007608 [06.05.2020].

Schroepfer, Mike, *Creating a data set and a challenge for deepfakes,* Facebook AI, 5. September 2019, https://ai.facebook.com/blog/deepfake-detection-challenge/ [06.05.2020].

Schwartz, Oscar, *You thought fake news was bad? Deep fakes is where news goes to die,* The Guardian, 12. November 2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth [06.05.2020].

Simonite, Tom, Forget Politics. *For Now, Deepfakes Are for Bullies.* Wired, 4. September 2019, https://www.wired.com/story/forget-politics-deepfakes-bullies/ [06.05.2020].

Soufan Center 2019, IntelBrief: *The Use of Disinformation in the British Election,* 13. Dezember 2019, https://thesoufancenter.org/intelbrief-the-use-of-disinformation-in-the-british-election [06.05.2020].

Stupp, Catherine, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Scams using artificial intelligence are a new challenge for companies,* Wall Street Journal, 30. August 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [06.05.2020].

T **Technopedia,** *Definition hashing*, https://www.techopedia.com/definition/14316/hashing [06.05.2020].

Techterms, *Definition hash*, https://techterms.com/definition/hash [06.05.2020].

Tolosana, Ruben / Vera-Rodriguez, Ruben / Fierrez, Julian / Morales, Aythami / Ortega-Garcia, Javier, *Deep-fakes and beyond: A survey of face manipulation and fake detection*, 2020, https://arxiv.org/abs/2001.00179 [06.05.2020].

V **Vincent, Brandi,** *Lawmakers Press Social Media Giants to Confront Deepfake Threats. Sens. Marco Rubio and Mark Warner want Facebook, YouTube, TikTok and others to create industry standards for handling synthetic content,* 2. Oktober 2019, https://www.nextgov.com/emerging-tech/2019/10/lawmakers-press-social-media-giants-confront-deepfake-threats/160325/ [06.05.2020].

Vosoughi, Soroush / Roy, Deb / Aral, Sinan, *The spread of true and false news online.* In: Science 359/2018, S. 1146–1151, https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf [06.05.2020].

W **Wagner, Travis Le / Blewer, Ashley,** *„The Word Real Is No Longer Real"*: Deepfakes, Gender, and the Challenges of AI-Altered Video. In: Open Information Science 3/2019, S. 32–46, https://www.researchgate.net/publication/334730810_The_Word_Real_Is_No_Longer_Real_Deepfakes_Gender_and_the_Challenges_of_AI-Altered_Video [06.05.2020].

Waterson, Jim, *Facebook refuses to delete fake Pelosi video spread by Trump supporters*, The Guardian, 24. Mai 2019, https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site [06.05.2020].

Weisberger, Mindy, *This Animated Mona Lisa Was Created by AI, and It Is Terrifying,* Live Science, 27. Mai 2019, https://www.livescience.com/65573-mona-lisa-deepfakes.html [06.05.2020].

Wu, Yue / Abdalmageed, Wael / Natarajan, Premkumar, *ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features.* In: IEEE Conference on Computer Vision and Pattern Recognition, 2019.

Y **YouTube,** *How YouTube supports elections*, 3. Februar 2020, https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html?m=1 [06.05.2020].

Rechtstexte:

B **Bundesregierung,** *Entwurf eines Gesetzes zur Änderung des Netzwerkdurchsetzungsgesetzes,* 31. März 2020, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RegE_Aenderung_NetzDG.pdf?__blob=publicationFile&v=2 [06.05.2020].

C ***Communications Decency Act of 1996,*** *Section 230, Pub. LA. No. 104–104, 110 Stat. 56*, 1996, https://transition.fcc.gov/Reports/tcom1996.pdf [06.05.2020].

G ***Genehmigungsgesetz zur nationalen Verteidigung (National Defense Authorization Act, NDAA) 2020,*** https://www.govinfo.gov/content/pkg/BILLS-116s1790enr/pdf/BILLS-116s1790enr.pdf [06.05.2020].

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken *(Netzwerkdurchsetzungsgesetz – NetzDG)*, https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html [06.05.2020].

K ***Kalifornisches Parlament, Gesetzesentwurf 730*** *Elections: deceptive audio or visual media,* https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 [06.05.2020].

Kongress der Vereinigten Staaten von Amerika, *Resolution 3230 des Repräsentantenhauses: Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019*, https://www.congress.gov/bill/116th-congress/house-bill/3230/text [06.05.2020].

Konrad-Adenauer-Stiftung und Counter Extremism Project

Dieser Bericht wurde in enger Zusammenarbeit mit der Konrad-Adenauer-Stiftung erstellt. Das Counter Extremism Project und die Konrad-Adenauer-Stiftung haben sich bei der Erstellung, Veröffentlichung und Platzierung dieses Berichts abgestimmt. Die Stiftung begleitet und fördert die Debatte über die Auswirkungen von Deepfakes auf unsere Gesellschaft in der deutschen Gesetzgebung und Öffentlichkeit.

Das Counter Extremism Project (CEP) ist eine gemeinnützige, überparteiliche, internationale Organisation, die das Ziel verfolgt, der Bedrohung durch extremistische Ideologien entgegenzuwirken und pluralistisch-demokratische Kräfte zu stärken. CEP befasst sich mit Extremismus in jeglicher Form – dazu gehören islamistischer Extremismus/Terrorismus als auch Rechts- und Linksextremismus/Terrorismus. CEP übt dazu auf Basis und mittels eigener Recherchen und Studien Druck auf finanzielle und materielle Unterstützungsnetzwerke extremistischer und terroristischer Organisationen aus, arbeitet den Narrativen von Extremisten und Terroristen sowie ihren Rekrutierungstaktiken im Internet entgegen, entwickelt bewährte Praktiken (*good practices*) zur Reintegration von Extremisten und Terroristen, und wirbt für effektive Regulierungen und Gesetze.

Neben Büros in den Vereinigten Staaten verfügt das CEP über ein Büro und eine eigenständige Rechtspersönlichkeit als Counter Extremism Project Germany gGmbH in Berlin und unterhält auch Vertretungen in Brüssel. Die Aktivitäten von CEP werden geleitet von einer internationalen Gruppe ehemaliger Politiker, leitender Regierungsbeamter und Diplomaten. CEP finanziert sich aus Spenden von Privatpersonen und projektbezogenen Fördermitteln. CEP unterstützt politische Entscheidungsträger auf der ganzen Welt bei der Ausarbeitung von Gesetzen und Vorschriften zur wirksamen Prävention und Bekämpfung von Extremismus und Terrorismus, insbesondere auch beim Kampf gegen die Finanzierung des Terrorismus.

Näheres dazu unter: www.counterextremism.com

Herausgeberin:
Konrad-Adenauer-Stiftung e. V. 2020, Berlin

Ansprechpartnerin in der Konrad-Adenauer-Stiftung:
Nauel Semaan
Referentin Terrorismusbekämpfung
nauel.semaan@kas.de

Umschlagfoto: © istock by Getty images/posteriori
Kapiteltrenner: S. 8, 17 © istock by Getty images/StudioM1; S. 11, 25 © istock by Getty images/a-r-t-i-s-t; S. 31 © alamy/Dmytro Razinko; S. 53 © istock by Getty images/Evgeny Gromov
Gestaltung und Satz: yellow too Pasiek Horntrich GbR
Übersetzung aus dem Englischen: Marie Liedtke, Redkey Translations

Die Printausgabe wurde bei der Druckerei Kern GmbH, Bexbach, klimaneutral produziert und auf FSC-zertifiziertem Papier gedruckt.
Printed in Germany.

Gedruckt mit finanzieller Unterstützung der Bundesrepublik Deutschland.

ISBN 978-3-95721-734-9

Die Verbreitung von Fake News als politisches Instrument ist längst Thema im politischen Diskurs. Dabei gilt es auch, auf technologische Neuerungen zu reagieren, die das Potenzial von Desinformationskampagnen fortlaufend ausbauen und somit unsere innere Sicherheit bedrohen. Diese Studie erläutert, wie das unregulierte Erstellen und die unkontrollierte Verbreitung von sogenannten Deepfakes – mit künstlicher Intelligenz veränderte Videos, die zur politischen Manipulation eingesetzt werden können – eine Bedrohung für unsere demokratische Gesellschaft darstellen. Die Autoren liefern außerdem Lösungsvorschläge für die Politik, um die Gefahr abzuwehren.

Konrad-Adenauer-Stiftung e. V.