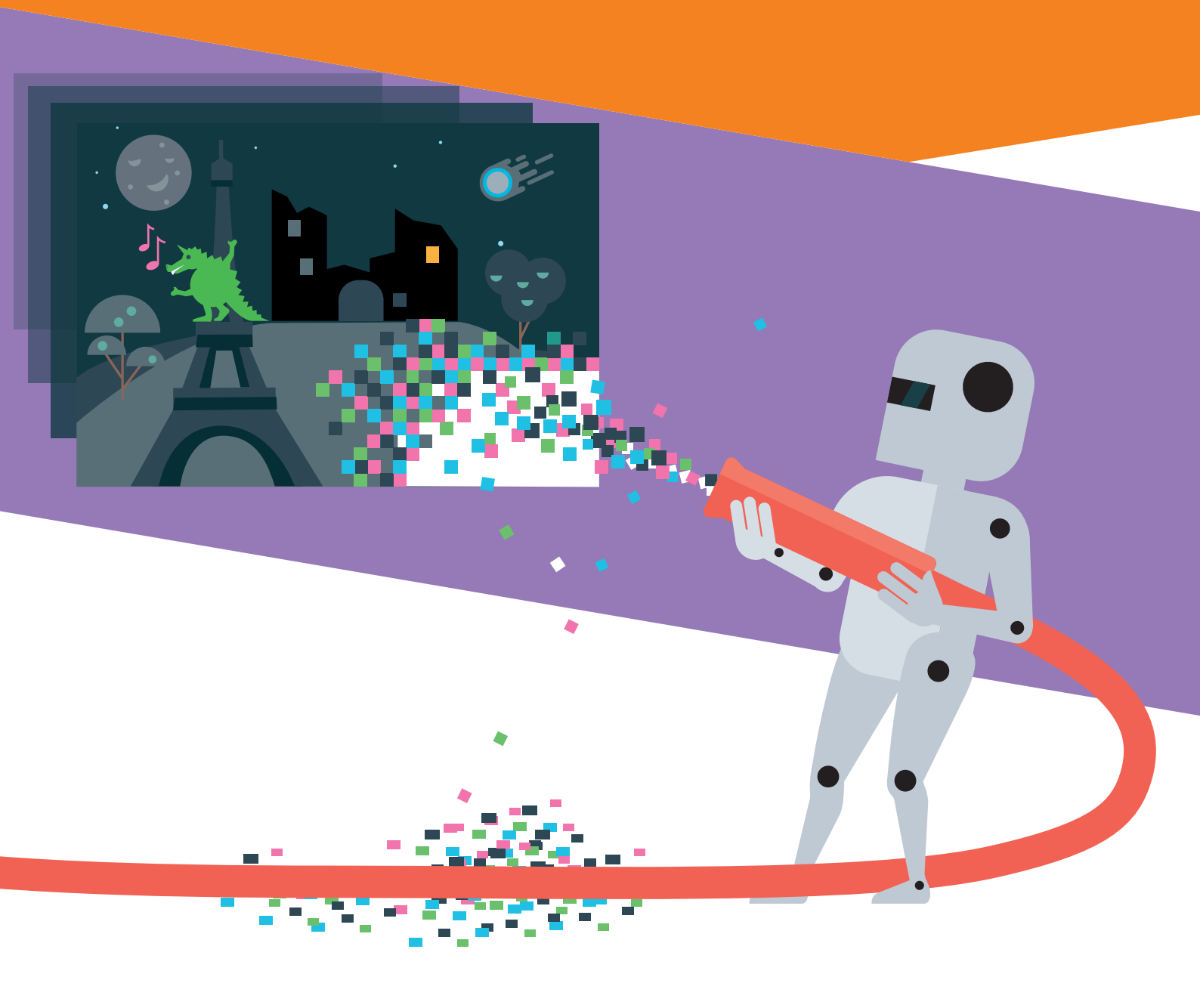
Augen und Ohren auf dem Prüfstand

Kurzfassung der Studie von TA-SWISS «Deepfakes und manipulierte Realitäten»



Die Stiftung TA-SWISS, ein Kompetenzzentrum der Akademien der Wissenschaften Schweiz, setzt sich mit den Chancen und Risiken neuer Technologien auseinander.

Die hier vorliegende Kurzfassung basiert auf einer wissenschaftlichen Studie, die im Auftrag von TA-SWISS von einem interdisziplinären Team unter der Leitung von Murat Karaboga (Fraunhofer-Institut für System- und Innovationsforschung ISI in Karlsruhe) durchgeführt wurde. Beteiligt waren zudem Nula Frei (Institut für Europarecht, Universität Freiburg i.Ue.), Manuel Puppis und Patric Raemy (Dept. für Kommunikationswissenschaft und Medienforschung, Universität Freiburg i.Ue.), Daniel Vogler (Forschungszentrum für Öffentlichkeit und Gesellschaft fög, Universität Zürich), Frank Ebbers (Competence Center Neue Technologie am Fraunhofer ISI, Karlsruhe), Greta Runge (Fraunhofer-Institut für System- und Innovationsforschung ISI in Karlsruhe), Adrian Rauchfleisch (Graduate Institute of Journalismus an der National Taiwan University), Gabriele de Seta (Department für Linguistik, Literatur and ästhetische Studien an der Universität Bergen), Gwendolyn Gurr (Audience Data Analyst, Schweizer Radio und Fernsehen SRF), Michael Friedewald (Fraunhofer-Institut für System- und Innovationsforschung ISI in Karlsruhe), Sophia Rovelli (Institut für Europarecht der Universität Freiburg i. Ue.).

Die Kurzfassung stellt die wichtigsten Resultate und Schlussfolgerungen in verdichteter Form dar und richtet sich an ein breites Publikum.

Kurzfassung der Studie von TA-SWISS «Deepfakes und manipulierte Realitäten – Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz»

Murat Karaboga, Nula Frei, Manuel Puppis, Daniel Vogler, Patric Raemy, Frank Ebbers, Greta Runge, Adrian Rauchfleisch, Gabriele de Seta, Gwendolyn Gurr, Michael Friedewald, Sophia Rovelli

TA-SWISS, Stiftung für Technologiefolgen-Abschätzung (Hrsg.).
vdf Hochschulverlag an der ETH Zürich, 2024.

ISBN 978-3-7281-4185-9

Die Studie steht als eBook zum freien Download bereit: www.vdf.ch

Die vorliegende Kurzfassung ist ebenfalls online verfügbar: www.ta-swiss.ch

Deepfakes in aller Kürze

Die Entwicklung der künstlichen Intelligenz (KI) schreitet rasend schnell voran. So auch die Herstellung «synthetischer» Filme, Bilder und Audioaufnahmen: Diese zeichnen keinen realen Sachverhalt auf, sondern werden von Computerprogrammen erzeugt. Da die Herstellung solcher Deepfakes immer einfacher wird, dürfte ihre gesellschaftliche Bedeutung rasch zunehmen. Chancen bieten sie in der Unterhaltungsbranche und in der Ausbildung und Schulung. Doch die Risiken – insbesondere in politischen Auseinandersetzungen, hinsichtlich des Mobbings von Individuen und mit Blick auf Wirtschaftsdelikte – sind nicht von der Hand zu weisen.

Deepfakes – oder synthetische Medien – nennt man Fotos, Videos oder Tonaufnahmen, die mittels künstlicher Intelligenz hergestellt werden und einen Sachverhalt zeigen, der sich in dieser Form nie ereignet hat. Es kann sich dabei um manipulierte Dateien handeln oder um solche, die zur Gänze künstlich sind, erzeugt von Software, die sich dabei auf Trainingsdaten aus riesigen Datenbeständen im Internet stützt. Die gegenwärtig etablierten Deepfake-Programme decken eine breite Spannweite ab, von einfach zu nutzender Software für den Austausch von Gesichtern bis zu anspruchsvollen Anwendungen für das «virtuelle Puppenspiel» mit künstlichen Personen. Zudem existieren bereits erste Programme, die – einstweilen noch rudimentäre – Videos aufgrund von Textbefehlen («Prompts») erzeugen können.

Einige Chancen …

Positive Potenziale bieten synthetische Medien für die Unterhaltungsindustrie. Auch andere wirtschaftliche Anwendungen sind vielversprechend – etwa, wenn künstliche Influencer Kleider oder andere Produkte vorführen. An den Schulen könnte der Geschichtsunterricht weiter an Attraktivität zulegen, wenn sich Avatare von Persönlichkeiten aus vergangenen Epochen – Cäsar, Katharina die Grosse, Napoleon – interaktiv mit den Schülerinnen und Schülern unterhielten. Auch Ermittlungsbehörden versprechen sich einen Nutzen aus der Möglichkeit, bei ihren kriminalistischen Untersuchungen Tathergänge zu visualisieren.

… und Risiken

Deepfakes können missbraucht werden, um Personen bei unstatthaften Handlungen zu zeigen, die sie nie begangen haben, oder um ihnen Worte in den Mund zu legen, die sie nie gesagt haben. Solche Videos oder Audioaufnahmen können dazu dienen, Menschen zu erpressen oder zu kompromittieren – ein Vorgehen, das in politischen Auseinandersetzungen zum Einsatz kommt. Auch in privaten Beziehungen kann es Verheerungen anrichten, etwa in Form gefakter Rachepornos.

In betrügerischer Absicht kann die geklonte Stimme einer Person verwendet werden, um an Geld aus deren Freundes- oder Familienkreis zu gelangen. Für andere Wirtschaftsdelikte, etwa für das Ausspähen von Geschäftsgeheimnissen, liessen sich die geklonten Stimmen von Vorgesetzten missbrauchen.

Für die Medien stellen sich erhebliche Herausforderungen, da sie Videos aufwendig verifizieren müssen, damit sie nicht selber zur Verbreitung von Deepfakes beitragen.

Vordringliche Empfehlungen

Angesichts der rasanten technischen Entwicklung kann nur eine Kombination verschiedener Schutzvorkehrungen sicherstellen, dass die positiven Potenziale synthetischer Medien ausgeschöpft und die schädlichen Auswirkungen begrenzt werden können.

Politische Massnahmen, Deepfake-Detektoren, die von grossen Software-Anbietern angedachte Kennzeichnung synthetischer Medien und die Sensibilisierung für Deepfakes durch die Medien müssen sich ergänzen. Wichtig ist zudem die Selbstverantwortung jeder und jedes Einzelnen: Internet-Videos sollten mit gesunder Skepsis betrachtet und private Bilder und Videos zurückhaltend hochgeladen werden.

Der Staat sollte gegenüber den grossen Online-Plattformen durchsetzen, dass für Personen nachteilige Deepfakes gelöscht werden. Da Einzelpersonen im Konflikt mit grossen Online-Plattformen meistens den Kürzeren ziehen, braucht es Fachstellen, welche die Opfer von Deepfakes – oder die Betroffenen ungerechtfertigter Löschungen – beraten und unterstützen. Opferberatungsstellen für Cyberdelikte sollten von Bund und Kantonen mit genügend Mitteln ausgestattet werden.

Verzerrter Blick auf die Wirklichkeit

Was wir mit unseren Augen sehen und mit unseren Ohren hören, halten wir in aller Regel für wahr. Insbesondere Videomaterialien werden als Abbild der Realität kaum angezweifelt. Zumindest war das bis vor wenigen Jahren der Fall. Denn heute ermöglicht es die Technik, mit geringem Aufwand täuschend echt aussehende Videos und Audioaufnahmen von Ereignissen herzustellen, die sich nie ereignet haben.

Im Spätherbst 2023 schien es für kurze Zeit, als stehe ein künftiges Supermodel kurz vor seiner Entdeckung: Ob der Videos, die Emily Pellegrini auf ihrem neuen Instagram-Kanal gepostet hatte, gerieten die Betrachter in Verzückung. Die Schar ihrer Follower wuchs rasant. Und zur langen Reihe an Herz- und Flammen-Emojis in der Kommentarspalte kamen etliche Kontaktanfragen hinzu. Zeitungen berichteten, ein deutscher Fussballer habe wiederholt um ein Date gebeten, und auch ein Milliardär, ein Tennisstar sowie andere Grössen aus der Sportwelt hätten sich um die Schöne bemüht – bis sie zur Kenntnis nehmen mussten, dass sie nicht um eine Frau aus Fleisch und Blut buhlten, sondern um einen mittels künstlicher Intelligenz (KI) erzeugten Avatar. Die «Traumfrau des Durchschnittsmannes» habe ihm als Vorbild gedient, gab ihr – anonym gebliebener – Schöpfer bekannt, dem die artifizielle Emily laut der britischen Daily Mail 10 000 Dollar pro Monat einbrachte.

Das nach eigenen Angaben «fun-loving girl» Emily Pellegrini, das seine äusserlichen Vorzüge auf zahlungspflichtigen Plattformen wie «Onlyfans» und «Fanvue» präsentiert, steht für eine neue Art von Influencern: Synthetisch hergestellte, aber täuschend echt wirkende Figuren vornehmlich weiblichen Geschlechts, die dank KI-basierter Textgeneratoren sogar mit ihrem Publikum chatten können. Die Vorteile, welche die künstlichen Personen der Werbung bringen, liegen auf der Hand: Einmal erschaffen, verlangen sie keinen Stundenlohn, werden niemals müde und befolgen jede Anweisung.

Tiefgreifende Täuschungen

Software, die auf künstlicher Intelligenz bzw. auf künstlichen neuronalen Netzwerken beruht, macht es möglich, Videos herzustellen, die einen Sachverhalt zeigen, der sich in Realität nie so zugetragen hat. Es kann sich dabei um Filme von Naturkatastrophen oder Explosionen handeln, die sich nie ereignet haben. Oder um Videos bekannter Persönlichkeiten, die etwas sagen oder tun, das sie nie gesagt oder getan haben. So zeigt ein Kurzfilm, hergestellt vom Videokünstler Bob de Jong aus Amsterdam, wie der ehemalige niederländische Ministerpräsidenten Mark Rutte mit tremolierendem Doppelkinn auf der Geige gefühlvoll den Solopart von «Stille Nacht»

intoniert. Hochgeladen hat Bob de Jong sein Werk auf seinem YouTube-Kanal «Diep Nep».

Übersetzt man «Diep Nep» auf Englisch, wird es zu «Deepfake». Diese Bezeichnung hat sich auch im deutschen Sprachraum etabliert, für authentisch scheinende Videos, die aber stark manipuliert oder zur Gänze am Computer erzeugt wurden. Die Fachwelt spricht denn auch von «synthetischen Medien». Der dazu erforderliche Rohstoff: Daten aus dem Internet, insbesondere Bilder, Videos und Tonaufnahmen aus den sozialen Medien und von Videoplattformen. Während allerdings Bob de Jong seine Schöpfungen offen als Artefakt deklariert, liegt bei zahlreichen – wenn nicht gar den meisten – Deepfakes die Urheberschaft im Dunkeln. In der vorliegenden Kurzfassung werden die Bezeichnungen «Deepfake» und «synthetische Medien» synonym verwendet.

Pionierwerke aus der Schmuddelecke

Die ersten Videofälschungen traten im Herbst 2017 auf Reddit in Erscheinung, einer Art elektronischem Sammelbecken für Inhalte aus sozialen Medien. Hochgeladen hatte die Filmchen ein User namens «DeepFake», der auf pornografischen Videos das Gesicht der originalen Darstellerin gegen das Antlitz von Emma Watson, Gal Gadot oder anderen Filmstars ausgetauscht hatte. Wenig später stellte ein anderer Reddit-User eine Software namens FakeApp zur Verfügung, die es allen ermöglichte, selber Deepfakes herzustellen. Was früher finanzkräftigen Hollywood-Studios vorbehalten war – die rechenintensive Produktion von 3D-Computergrafiken – war somit für alle realisierbar geworden, die es schafften, die fünf von FakeApp vorgeschriebenen einfachen Herstellungsschritte auszuführen.

In der Folge wurden im Darkweb massenhaft manipulierte Pornos hochgeladen. Die Videos, zunächst aufgrund ihrer geringen Auflösung und der ruckartigen Bewegungen leicht als Fälschungen zu entlarven, wirkten dank technischer Fortschritte immer echter. Längst werden nicht mehr nur die Bilder Prominenter für solche Machwerke missbraucht. Vielmehr können heute alle zum Opfer werden, die jemanden verärgert haben; der Ausdruck «Racheporno» erhielt im Februar 2018 einen eigenen Eintrag auf Wikipedia, der auch die Herstellung gefakter Nacktfilme erwähnt.

Schätzungen zufolge zeigt auch heute ein Grossteil der hochgeladenen Deepfakes pornografische Darstellungen von Frauen – obschon das Genre mittlerweile in andere Sparten, insbesondere in die Politik, hinübergeschwappt ist. Denn Persönlichkeiten, von denen zahlreiche Aufnahmen im Netz kursieren, liefern besonders viel Material, das für die Herstellung der Videotäuschungen genutzt werden kann.

Im Wettstreit zum besten Fake

Eine Technik, die bei der Erzeugung von Deepfakes viel Aufmerksamkeit auf sich zieht, heisst Generative Adversarial Networks, kurz: GAN. So nennt man Computerprogramme, die im Vergleich zu einem Trainingsset ähnliche, aber neuartige Bilder erzeugen können. Die Programme bestehen aus zwei Teilen, dem Generator und dem Diskriminator, die gegeneinander antreten. Während der Generator ähnliche Bilder wie jene aus dem Trainingsset zu erzeugen sucht, ist der Diskriminator darauf ausgerichtet, Unterschiede zwischen den neu erzeugten Bildern und den Trainingsdaten auszumachen. Im Wettlauf gegeneinander verbessern sich Generator und Diskriminator gegenseitig, sodass am Ende Bilder entstehen, die sich stilistisch kaum mehr von den echten Vorlagen unterscheiden lassen.

Der alternative Ansatz besteht im Einsatz von Autoencodern. Das sind künstliche neuronale Netze, die in der Lage sind, wesentliche Merkmale aus einem Bilddatensatz herauszufiltern und auf andere Bilder zu übertragen.

Puppenspiel mit fremdem Körper

Die bei Deepfakes vorgenommenen Manipulationen greifen unterschiedlich tief in die Trickkiste – was indes nichts über die jeweilige Wirkung auf das Publikum aussagt.

So können sich die Eingriffe darauf beschränken, Gesichtsausdruck und Mundbewegungen einer Person zu verändern. Dabei wird die Mimik eines Schauspielers oder einer Schauspielerin auf die Zielperson übertragen; in der Fachsprache nennt man diese Neuinszenierung eines Gesichts «facial reenactment». Dieser Ansatz kommt auch zum Zug, wenn sich eine Person auf einem Video selber synchronisiert, d.h., wenn ihre Mundbewegungen auf ihre Aussagen in einer anderen Sprache abgestimmt werden. So hat etwa die Firma HeyGen Labs ein KI-gestütztes Videoprogramm entwickelt, das den Ton aus einem Video aufzeichnet, den Inhalt übersetzt und das Gesagte in der anderen Sprache wieder ins Video überträgt – und zwar so, dass die

Stimme der Sprecherin oder des Sprechers geklont wird und sich zugleich die Lippen passend zum Gesprochenen bewegen. So können beispielsweise Journalisten und Moderatorinnen ihre Wortbeiträge gleich selber synchronisieren.

Eine andere Art von Deepfakes, das sog. Gesichts-Morphing, besteht darin, die Züge zweier Individuen miteinander zu verschmelzen. Diese Technik wird in erster Linie in kriminellen Milieus eingesetzt, um Ausweisdokumente so zu fälschen, dass sie von mehreren Personen zugleich verwendet werden können.

Der Gesichtsaustausch, das «face swapping», kam in der eingangs geschilderten frühen Phase der Deepfakes zum Einsatz, bei der Produktion gefälschter Pornovideos. Diese Technik wird auch spielerisch genutzt. Im Internet kursieren kostenlose Apps, um ikonische Filmszenen «nachzudrehen», indem etwa das Gesicht von Leonardo di Caprio oder Kate Winslet durch das eigene Porträt ersetzt wird.

KI-gestützte Software ist zudem in der Lage, von Grund auf neue Bilder von Menschen zu konstruieren, die es gar nicht gibt. Solche von Gesichtsgeneratoren erzeugten Porträts und Videos finden als Avatare in Videospielen Verwendung, oder als virtuelle Gesprächspartner im vollautomatisierten Kundendienst.

Wie ein Puppenspieler agiert die Software schliesslich dann, wenn sie die Posen und Bewegungen eines ganzen Menschen in einem Video verändert. Diese Art von Deepfakes, «full body puppetry» genannt, gilt als die komplizierteste. Auch gibt es derzeit noch kein KI-Gesamtpaket, das imstande wäre, ein komplettes Video mit Sprachanimation und Stimmsynthetisierung zu erstellen. Vielmehr müssen mehrere zum Teil kostenpflichtige und kompliziert zu bedienende Programme miteinander verbunden werden, was die Erstellung echt wirkender Deepfake-Videos zu einer schwierigen Angelegenheit macht.

Die Herstellung synthetischer Videos auf Knopfdruck dürfte allerdings bald in Griffnähe rücken: Nach dem Vorbild KI-gestützter Bildgeneratoren, die aufgrund von Textbefehlen die fotorealistische Aufnahme eines Einhorns oder ein Falsifikat einer Malerei von Rembrandt erzeugen, ermöglichen es erste KI-gestützte Programme, anhand von Sprach- oder Textbefehlen Deepfake-Videos zu generieren. Solche Software dürfte in absehbarer Zeit für breitere Anwenderkreise zugänglich werden.

Stimmen aus der Retorte

Nicht nur das Auge kann getäuscht werden, sondern auch das Ohr – durch Software, die in der Lage ist, Stimme und Sprechgewohnheiten eines Menschen zu klonen. In Fachkreisen Berühmtheit erlangte 2018 das Experiment einer schottischen Firma, die es fertigbrachte, den ermordeten US-Präsidenten John F. Kennedy zumindest akustisch zu reanimieren: Aufgrund des Manuskripts und zahlreicher Sprachaufzeichnungen des Staatsmannes gelang es, eine Audiowiedergabe der Rede zu erzeugen, die dieser im Herbst 1963 in Dallas nicht hatte halten können, weil er einem Attentat zum Opfer gefallen war. Sowohl Bostoner Akzent als auch Sprechkadenz Kennedys wurden perfekt imitiert.

In den letzten Jahren ist die Technik weiter fortgeschritten. Um ein Modell der Sprechweise eines Menschen zu erzeugen, genügen ein Standard-Laptop und wenige Sekunden eines Audioclips – beispielsweise aus einem Vortrag, der auf YouTube hochgeladen wurde. Fortschritte hat auch die sogenannte Sprachsynthese gemacht. Diese kommt in Software zur Anwendung, die Geschriebenes in Audioclips umwandelt. Sie dient zur automatischen Herstellung von Hörbüchern, und auch Sehbehinderte nutzen sie, wenn sie sich Texte vorlesen lassen.

Instrumente zur Identifikation von Deepfakes

Allein schon der alarmierende Ausdruck «Deepfakes» weist auf das Ziel hin, das viele synthetische Videos und Tonaufnahmen verfolgen: Sie sollen das Publikum in die Irre führen und zugunsten des Absenders beeinflussen. Die Fachwelt diskutiert verschiedene Möglichkeiten des Umgangs mit visuellen und akustischen Täuschungen.

Ein Ansatz besteht darin, den Ursprung einer Aufnahme transparent zu machen. Dies könnte dank einer digitalen Signatur geschehen, die direkt während der Aufnahme auf einer Video- oder Audiodatei angebracht würde. Basierend auf der Blockchain wäre dies technisch relativ leicht umzusetzen; mit viel Aufwand liesse sich ein solcher «digitaler Fingerabdruck» allerdings fälschen. Kritische Stimmen warnen zudem davor, autoritären Regimes und Geheimdiensten würde damit ein Instrument in die Hand gegeben, um Whistleblower, Menschenrechtsaktivisten und unliebsame Journalistinnen aufzuspüren. Des Weiteren könnte eine Signatur zwar den Ursprung einer Aufnahme bezeugen – nicht aber, ob tatsächlich das ganze Geschehen oder nur ein Ausschnitt daraus aufgenommen wurde. Auch Videos, die mit Schauspielerinnen und Schauspielern nachgestellte Ereignisse zeigten, liessen sich dadurch nicht ermitteln. Solche Videos wären zwar technisch «echt», könnten aber dennoch Unwahrheiten verbreiten. Ein solches Szenario greift eine norwegische Fernsehserie auf, in der eine Partisanengruppe ein Video der – angeblichen – Ermordung eines umstrittenen Ministerpräsidenten zeigt, damit dieser besser untertauchen kann.

Fälschungsmerkmale detektieren

Andere Methoden zielen darauf ab, artifizielle Videos oder Audioaufnahmen anhand bestimmter Merkmale zu demaskieren. Doch mit Blick auf geklonte Stimmen herrscht weitgehend Einigkeit: Diese akustischen Deepfakes klingen mittlerweile dermassen echt, dass sie sich – insbesondere, wenn sie am Telefon abgespielt werden – kaum von der Stimme einer echten Person unterscheiden lassen. Polizeibehörden warnen denn auch vor dem «Enkeltrick 3.0», der umso besser funktionieren dürfte, als das Opfer die Stimme eines Familienmitglieds zu hören glaubt, das um Geld bittet.

Videofälschungen wiederum können sich durch bestimmte Artefakte verraten, die zwar von menschlichen Augen nicht unbedingt erkannt werden, wohl aber von KI-gestützten Algorithmen. Fehlerhafte Kanten, unnatürliche Überblendungen, Verformungen und Unschärfen können auf einen Deepfake hinweisen. Mittlerweile gibt es Detektor-Programme, die für sich in Anspruch nehmen, gefakte Videos zu

entlarven. Im Rahmen der Studie von TA-SWISS wurden zwei kostenlose Detektoren getestet. Die Ergebnisse vermochten nicht zu befriedigen, denn beide Programme lieferten falsche Ergebnisse. Problematisch war nicht zuletzt, dass die Detektoren einige echte Videos als Fälschungen deklarierten – was die Vertrauenswürdigkeit originaler Inhalte untergraben könnte. Ohnehin ist zu erwarten, dass Entwicklerinnen und Entwickler von Deepfake-Software die verräterischen Merkmale ebenfalls kennen und alles daran setzen, ihre Programme entsprechend zu verbessern – ein Katz- und Maus-Spiel, bei dem auf absehbare Zeit die Fälscherinnen und Fälscher die Nase vorn haben dürften.

Gesundes Misstrauen beibehalten

Zurzeit dürfte gesunder Menschenverstand den technischen Detektoren ebenbürtig sein, wenn es darum geht, Deepfakes zu erkennen. Kritisches Hinterfragen der Quelle und der Inhalte eines Videos, Wachsamkeit gegenüber Unstimmigkeiten von Details wie Haarsträhnen, Fingern oder Ohrringen sowie Aufmerksamkeit gegenüber einem ungewöhnlichen Verhalten der gefilmten Person können helfen, Deepfakes zu entlarven. Ausserdem lässt sich die Fähigkeit, Videofälschungen zu erkennen, auf Webseiten wie Detectfakes trainieren.

Etwas mehr Argwohn hätte es jedenfalls dem deutschen Fussballer und den anderen Bewunderern von Emily Pellegrini erspart, ihr auf den Leim zu gehen. Beim Vergleich verschiedener Videos ist festzustellen, dass die Proportionen der künstlichen Influencerin variieren. Auch ihre durchwegs gut gelaunten bis zuweilen leicht unterwürfigen Antworten in den Chats sollten misstrauisch stimmen. Jedenfalls ist bei surrealer Perfektion und übertriebenem Glamour Skepsis geboten.

Deepfakes mit einem variantenreichen Methodenset ausgeleuchtet

In der Studie von TA-SWISS über Deepfakes wurden nebst einer gründlichen Literaturrecherche mehrere Befragungen durchgeführt. Mittels einer Online-Befragung, ergänzt durch ein Online-Experiment, wurde die Bevölkerung nach ihren Erfahrungen und ihrem Umgang mit Deepfakes gefragt. Zudem erkundigte sich die Projektgruppe in mehreren Interviews und Befragungen bei Medienschaffenden, bei Verwaltungsangestellten sowie bei Politikerinnen und Politikern nach ihrer Einschätzung der Risiken, Chancen und Auswirkungen von Video- und Tonfälschungen. Schliesslich testete die Projektgruppe mehrere kostenlose Deepfake-Detektoren auf ihre Erkennungsfähigkeit von Videofälschungen.

Wie die Bevölkerung und Medienschaffende Deepfakes wahrnehmen

In der Schweiz ist die Bevölkerung bis jetzt noch kaum mit Deepfakes in Berührung gekommen. Am ehesten begegnet man ihnen auf Plattformen wie YouTube, TikTok und Instagram. Die im Rahmen der Studie Befragten bringen Deepfakes überwiegend mit Risiken in Verbindung und sind kaum in der Lage, gut gemachte Deepfake-Videos von realen Videos zu unterscheiden. Auch in den grössten Schweizer Medienhäusern nimmt man Deepfakes vor allem als Risiko wahr. Mit der Thematisierung von Deepfakes kommt journalistischen Medien eine wichtige Rolle bei der Sensibilisierung der Bevölkerung zu.

Die Studie von TA-SWISS ist die erste gesamtheitliche Untersuchung, die sich mit der Wahrnehmung von Deepfakes in der Schweiz befasst. Von über 1300 Befragten erklärte etwas mehr als die Hälfte, den Ausdruck «Deepfake» zu kennen – und etwas weniger als die Hälfte, schon einmal ein solches Video gesehen zu haben. Eigene Erfahrungen im Erstellen oder Verbreiten von Deepfakes hat hierzulande eine kleine Minderheit von zwei bzw. drei Prozent. Die Resultate aus der Studie von TA-SWISS zeigen insgesamt, dass die Menschen in der Schweiz eher wenig Erfahrungen mit Deepfake-Technologien haben. Typische Einflussgrössen wie Alter, Geschlecht und Bildung, die in der Regel bei der Aneignung neuer Techniken ins Gewicht fallen, haben hier keinen allzu starken Einfluss.

Gesellschaft stärker bedroht als das Individuum

Die Schweizer Bevölkerung nimmt Deepfake-Technologien eher als Risiko denn als Chance wahr. In erster Linie wird befürchtet, als Deepfakes verbreitete Falschnachrichten könnten das Vertrauen in die Schweizer Informationsmedien untergraben. Als etwas weniger virulent wird die Gefahr eingeschätzt, Deepfakes könnten in der Schweiz Abstimmungen oder Wahlen beeinflussen.

Das Risiko, persönlich ein Opfer von Deepfakes zu werden, schätzen die Befragten als relativ gering ein. Augenfällig ist, dass Frauen diese Gefahr höher einstufen als Männer – ein Ergebnis, das angesichts der vielen pornografischen Deepfakes wenig überrascht.

Das Label beeinflusst die Wahrnehmung von Chancen

Erkundigt man sich nach allfälligen Chancen, die mit Deepfakes einhergehen könnten, äussern sich die Befragten skeptisch. Dies ändert sich, wenn statt der Bezeichnung «Deepfake» der neutralere Ausdruck «synthetische Medien» verwendet wird. In einer Vorstudie wurden die Befragten in zwei Gruppen aufgeteilt, die einen unterschiedlichen Fragebogen erhielten. Die eine Gruppe erhielt den Fragebogen mit dem Begriff «Deepfake». Für die zweite Gruppe wurde für alle Fragen die Bezeichnung «synthetische Medien» verwendet.

Es zeigte sich, dass der Ausdruck «synthetische Medien» weniger bekannt ist: Während rund zwei Drittel der Befragten mit der Bezeichnung «Deepfakes» etwas anfangen können, sind es bei «synthetische Medien» bloss etwas über einem Drittel. Die Risikoeinschätzung fällt bei beiden Labels etwa gleich aus. Anders verhält es sich bei den Chancen: Synthetischen Medien werden im Hinblick auf ihre Wirkung in den Medien und in der Wirtschaft deutlich höhere Chancen zugestanden als Deepfakes. Wie die Gesellschaft den Nutzen einer Technik wahrnimmt, hängt also nicht zuletzt vom Etikett ab, das an dieser haftet.

Tipps nützen wenig, Vertrautheit mit neuen Medien hilft

In der Studie von TA-SWISS zeigte sich, wie schwer es fällt, Deepfakes zu erkennen. In einem Experiment wurden den Befragten drei Deepfake- und drei reale Videos abgespielt, deren Realitätsgehalt sie einschätzen sollten. Für dieses Experiment wurden die Befragten wiederum zwei Gruppen zugeteilt, von denen die eine zuvor eine kurze Hilfestellung erhielt, um Deepfakes zu erkennen.

Das Fazit: Die Befragten waren sich in ihrer Einschätzung sehr unsicher. Sie sind folglich kaum in der Lage, gut gemachte Deepfake-Videos von realen Videos zu unterscheiden. Zudem konnte die Gruppe, der zuvor Tipps für das Erkennen von Deepfakes gegeben wurde, die gezeigten Videos nicht besser einschätzen als die Gruppe, die keine Hilfestellung erhalten hatte.

Hingegen stellte sich heraus, dass Erfahrungen im Umgang mit sozialen Medien positiv mit dem Erkennen von Deepfakes korrelieren. Mithin trägt übergeordnete Medienkompetenz dazu bei, Videofälschungen nicht aufzusitzen: Man sollte also nicht nur mit herkömmlichen Massenmedien vertraut sein, sondern auch den vorsichtigen Umgang mit Informationen aus unbekannten Quellen auf sozialen Medien erlernen.

Herausforderungen für den Journalismus

Falsche Angaben und Desinformation zu erkennen, gehört zum Kernauftrag von Medienschaffenden. Täuschend echt aussehende Videos und Audioaufnahmen stellen diese dabei vor zusätzliche Herausforderungen. Da Journalismus das politische Geschehen kritisch begleiten und zur Meinungs- und Willensbildung in der Öffentlichkeit beitragen soll, ist die korrekte Erkennung von Deepfakes durch Medienschaffende für die Gesellschaft insgesamt relevant. Deepfakes nicht (unbeabsichtigt) weiterzuverbreiten, liegt aber auch im eigenen Interesse der Medien: Denn ihrer Glaubwürdigkeit – und damit dem Geschäftsmodell der Medienhäuser – droht ansonsten ein schwerwiegender Reputationsschaden.

Medienschaffende stehen vor der Aufgabe, in möglichst kurzer Zeit die Echtheit von Videos (bzw. Audioaufnahmen) zu verifizieren. Doch solche Überprüfungen sind aufwendig. Zugleich stehen viele Medienhäuser unter finanziellem Druck; nicht alle können sich spezialisiertes Personal hierfür leisten. Ausserdem vermögen journalistische Beiträge die Öffentlichkeit zwar für Deepfakes zu sensibilisieren. Wenn diesen in der Berichterstattung aber (zu) viel Raum zugestanden wird, könnte dies die Gefahr erhöhen, dass sich in der Bevölkerung übertriebene Skepsis verbreitet und generell das Misstrauen gegenüber Medieninhalten zunimmt.

Journalistinnen und Journalisten exponieren sich in der Öffentlichkeit stark. Wie Erfahrungen aus Indien und den USA zeigen, können prominente Medienschaffende selber Opfer von Deepfakes werden. Und obgleich Schweizer Journalistinnen und Journalisten in der Regel nicht so prominent sind wie manche ihrer ausländischen Kolleginnen und Kollegen, werden auch hierzulande viele bedroht. Deepfakes könnten das Arsenal der Einschüchterungen erweitern.

Gedämpfter Alarm in Schweizer Redaktionsstuben

Die im Rahmen der Studie von TA-SWISS bei Schweizer Medienschaffenden durchgeführte Befragung zeigt, dass das Phänomen der Deepfakes auf den Redaktionen zwar wahrgenommen und in der journalistischen Ausbildung behandelt wird, aber nicht als dringliches Risiko gilt. Vielmehr schätzt man die Videofälschungen als eine Unterkategorie von Desinformation ein. Medienschaffende in der Schweiz befürchten derzeit (noch) nicht, selber Opfer von Deepfakes zu werden.

Mit Videofälschungen konfrontiert werden hiesige Redaktionen vor allem bei der Auslandsberichterstattung, etwa im Zusammenhang mit dem Krieg in der Ukraine. Hier sind die Redaktionen gefordert, gefälschte Videos zu erkennen, um sie nicht unwissentlich weiter zu verbreiten. Diesbezüglich könnten hiesige Redaktionen davon profitieren, dass grosse ausländische Medien mit gut dotierten Rechercheteams die Echtheit von Videos überprüfen würden, so ein Ergebnis aus der Befragung. Die Schweiz hingegen stehe nicht im Fadenkreuz von Deepfake-Produktionsstätten, weil Schweizer Medien international weniger Beachtung fänden.

Die Befragung in den Schweizer Medienhäusern zeigte aber auch, dass insbesondere komplexe Fälle hohe Ansprüche an die Verifikation stellen – und dass sorgfältige Überprüfungen in den Redaktionen nicht genügten, sondern durch Aufklärung und Sensibilisierung der Öffentlichkeit ergänzt werden müssten. Denn es reicht nicht, wenn die Medien den Wahrheitsgehalt von Neuigkeiten kritisch überprüfen. Vielmehr braucht es in der ganzen Gesellschaft ein Bewusstsein für manipulierte Information.

Verlässliche Quellen als Gegensatz zu Deepfake-Schleudern

Unter den befragten Medienschaffenden gelten Deepfakes überwiegend als Risiko. Ein gewisses Potenzial wird allenfalls in der Personalisierung von News-Angeboten durch ‹synthetische› Moderation erkannt, oder durch den Einsatz von Avataren bei Recherchen.

Der einzige Nutzen, der gefälschten Videos und Bildern zugestanden wird: Sie dürften dazu beitragen, die Position journalistischer Medien als zuverlässige Informationsquelle zu stärken – sofern es diesen gelingt, Deepfakes und andere Manipulationen frühzeitig zu erkennen und sich so von weniger vertrauenswürdigen Quellen abzuheben.

Unterschiedliche rechtliche Ansprüche an journalistische Medien und Online-Plattformen

Mit Blick auf die Verbreitung gefälschter Videos fallen die unterschiedlichen rechtlichen Vorgaben für journalistische Medien und Online-Plattformen ins Gewicht. Die traditionellen Medien finden in der Bundesverfassung der Schweizerischen Eidgenossenschaft Erwähnung, indem etwa Artikel 93 vorschreibt, dass Radio und Fernsehen die Ereignisse sachgerecht darstellen und die Vielfalt der Ansichten angemessen zum Ausdruck zu bringen müssen. Auch das Bundesgesetz über Radio und Fernsehen hält bei den Mindestanforderungen an den Programminhalt fest, Tatsachen und Ereignisse müssten sachgerecht dargestellt sein, «so dass sich das Publikum eine eigene Meinung bilden kann. Ansichten und Kommentare müssen als solche erkennbar sein.» Ausserdem gibt es Instanzen, bei denen Beschwerden eingereicht werden können, sollten Medien das Gebot der Sachgerechtigkeit und Ausgewogenheit verletzen.

Dagegen müssen Online-Plattformen wie soziale Netzwerke oder Video-Sharing-Dienste, die Inhalte ihrer User verbreiten, nicht darauf hinweisen, ob es sich bei einem Video um ein Deepfake handelt. Vielmehr sind die auf sozialen Medien verbreiteten Inhalte durch die Meinungsfreiheit geschützt, sodass der Staat nur gegen offensichtlich rechtswidrige Inhalte vorgehen kann. Bei pornografischen Videos kann zudem das Bundesgesetz über den Jugendschutz in den Bereichen Film und Videospiele zur Anwendung kommen. Das Gesetz verpflichtet die Verbreiter solcher Inhalte – darunter auch Streaming-Dienste –, Massnahmen zum Schutz Jugendlicher zu ergreifen und gegebenenfalls den Zugang zu entsprechenden Videos zu beschränken. Die Durchsetzung Schweizer Rechts gegenüber Angeboten aus dem Ausland gehört allerdings zu den grössten Herausforderungen im Umgang mit Deepfakes.

Wenn Avatare Politik machen und die Wirtschaft aufmischen

Ob bei kriegerischen Auseinandersetzungen oder im Wahlkampf: Wenn es hart auf hart geht, bedient man sich der Deepfakes, um die Gegenseite zu verunsichern. In der Wirtschaft fügen sich die gefakten Filme ins Repertoire der Cyberkriminalität ein.

Wenig überraschend, kommen Deepfakes im Wahlkampf zum Einsatz, um politische Gegner zu diskreditieren und das Zielpublikum zu verwirren.

So zeigte ein im März 2022 verbreiteter Deepfake den ukrainischen Präsidenten Selenskyj, wie er in einer Rede angeblich die Bevölkerung seines Landes zur Kapitulation vor den russischen Streitkräften auffordert. In Pakistan wandte sich Anfang 2024 der inhaftierte Oppositionsführer Imran Khan aus dem Gefängnis an seine Landsleute – und griff mit einem KI-Klon seiner selbst in den Wahlkampf ein. Auch in den USA wurde versucht, den Wahlkampf mittels Deepfake-Tricksereien zu beeinflussen. So erhielten Angehörige der demokratischen Partei in New Hampshire im Januar 2024 ein gefälschtes Telefonat von Joe Biden. In diesem «Robo-Call» forderte der Präsident die Angerufenen auf, den Vorwahlen fern zu bleiben.

In der Schweiz traten synthetische Medien in politischen Auseinandersetzungen ebenfalls bereits in Erscheinung. Im Sommer 2023 sorgte ein gänzlich durch KI generiertes Wahlplakat für Aufruhr, das eine durch Klimaaktivisten behinderte Ambulanz prominent ins Bild setzte – ein Vorfall, der sich in dieser Form hierzulande nie ereignet hat. Im Oktober des gleichen Jahres zeigte ein gefaktes Video Nationalrätin Sibel Arslan bei einem Aufruf, der im Widerspruch zu ihren erklärten Werten und Anliegen stand.

Humor als Wahlkampfhelfer

Fakt ist: Deepfakes können in der Öffentlichkeit Verwirrung stiften und politisch aktive Persönlichkeiten in Verruf bringen, einschüchtern oder ihnen allenfalls vertrauliche Informationen entlocken. Doch ungeachtet ihrer zersetzenden Wirkung verfügen synthetisch erzeugte Videos auch über positives politisches Potenzial. Mit humorvollen Beiträgen könnten sie die politische Debatte anregen und die Meinungsbildung unterstützen. Und bei Abstimmungen liessen sich Deepfakes einsetzen, um komplexe Sachverhalte zu veranschaulichen.

Politikerinnen und Politiker könnten sich überdies mit transparent gekennzeichneten, satirischen und humoristischen Deepfakes an ihre Wählerschaft wenden. Denn Humor und Witz eignen sich bestens, um die eigene Reichweite zu vergrössern und die Aufmerksamkeit der Bevölkerung auf sich zu ziehen.

Mehr Wachsamkeit im politischen Betrieb erwünscht

Die Autorinnen und Autoren der Studie von TA-SWISS erkundigten sich bei Mitgliedern des Schweizer Parlaments und bei Mitarbeitenden der eidgenössischen Verwaltung, wie sie Deepfakes wahrnehmen und einschätzen.

Im politischen Alltag sind die gefakten Videos angekommen. Denn eine Mehrheit der Befragten gab an, Deepfakes seien bei ihrer Arbeit bereits ein Thema. Die Antwort auf die Frage, ob die Videofälschungen eher als Risiko oder allenfalls als Chance zu beurteilen seien, fiel einhellig aus: Positive Aspekte vermochte nahezu niemand zu erkennen, die Befragten sahen ausschliesslich Risiken.

Im Vordergrund standen dabei die Gefahren für die Schweizer Demokratie und für das Vertrauen in die hiesigen Institutionen. Dass sie selber Opfer bzw. Protagonisten eines Deepfakes werden bzw. auf einen hereinfallen könnten oder dass ein gefälschtes Video internationale Beziehungen trüben könnte, wurde von den Befragten ebenfalls als relevante Risiken genannt – wenngleich sie der Ansicht waren, solche Ereignisse seien eher unwahrscheinlich. Einig waren sich die Befragten zudem in ihrer Einschätzung, es würden noch zu selten konkrete Schutzmassnahmen gegen Deepfakes ergriffen.

Potenziale für Unterhaltung und Bildung

Weniger negativ als in Politik und Verwaltung treten Deepfakes in der Wirtschaft in Erscheinung. Die Unterhaltungsindustrie gesteht ihnen einen vielfäl-

tigen Nutzen zu – etwa in der Filmindustrie. In der Gaming-Szene wird nach Möglichkeiten gesucht, die Gesichter der Spielerinnen und Spieler auf ihre Avatare zu übertragen. Und die Werbebranche verspricht sich Vorteile von künstlichen Influencern, die Kleider präsentieren können oder in der Unternehmenskommunikation aktiv sind.

Deepfakes können auch in Kampagnen gemeinnütziger Institutionen beim Einwerben von Spenden helfen. Ein synthetisches Double des ehemaligen Fussballstars David Beckham rief 2019 in neun verschiedenen Sprachen dazu auf, eine Petition zum Kampf gegen Malaria zu unterzeichnen und damit führende Köpfe aus den von der Krankheit besonders betroffenen Ländern zu mehr Engagement gegen die Krankheit aufzufordern.

In Schulstuben könnte sich das Interesse am Geschichtsunterricht erhöhen, wenn sich Kleopatra, Napoleon oder andere historische Persönlichkeiten als Deepfakes interaktiv mit den Jugendlichen unterhalten würden. Avatare könnten zudem im personalisierten Fernunterricht die Lernmotivation junger Menschen steigern. Auch in der Medizin werden synthetischen Medien positive Potenziale zugeschrieben, etwa bei der Therapie von Angststörungen: Dabei wird ein Avatar der behandelten Person in eine für diese normalerweise angsterregende Lage versetzt wie etwa das Balancieren in grosser Höhe. Aus psychologischer Sicht gestattet dies eine objektive Auseinandersetzung mit der Situation.

Wirtschaftsspionage mit geklauter Identität

Doch ohne Risiken sind Deepfakes auch für die Wirtschaft nicht – können sie hier doch ähnliche Reputationsschäden wie in der Politik anrichten: Gefälschte private Aussagen eines vermeintlichen Insiders vermögen den Ruf eines Unternehmens zu ruinieren oder die Aktienmärkte zu manipulieren. Gefakte Influencer wiederum können für Werbebetrug eingesetzt werden und untergraben letztlich das Vertrauen in den Anbieter.

Darüber hinaus erleichtern Deepfakes den Identitätsbetrug. Mit einer geklonten Stimme bzw. dem dreidimensionalen Avatar eines Menschen lassen sich Stimm- oder Gesichtserkennungssysteme überlisten. So kommen Kriminelle an das Konto einer Privatperson, oder sie können Geschäftsgeheimnisse ausspähen.

Freilich sind Deepfake-Angriffe auf Akteure in der Wirtschaft nichts Neues, sondern fügen sich in das Repertoire herkömmlicher Cyberkriminalität ein. Auch die Ziele von Cyber- und Deepfake-Delinquenten ähneln sich: Im Vordergrund stehen finanzielle Interessen oder die Sabotage der Konkurrenz. Besonders gefährdet sind Firmen oder Personen, die vier Kriterien erfüllen: Sie haben einen grossen Wert, ihre Sichtbarkeit ist hoch, in ihren Aktionen sind sie eher träge, und der Zugang zu ihnen ist relativ einfach.

Die Schweiz als attraktives Ziel

Als eine der innovativsten und produktivsten Wirtschaften weltweit gilt die Schweiz als attraktives Ziel für Cyber- und damit auch Deepfake-Angriffe – umso mehr, als ihre Sicherheitsarchitektur hinter ihrer bedeutenden wirtschaftlichen Rolle her hinkt: Dem Global Cybersecurity Index 2020 zufolge belegt sie Rang 42 von 182 Staaten. In einer Studie aus dem gleichen Jahr, ausgeführt im Auftrag des Nachrichtendienstes des Bundes, gaben 15 Prozent der befragten Unternehmen an, bereits von Wirtschaftsspionage betroffen gewesen zu sein. Allerdings meldete nur eine Minderheit – nämlich 13 Prozent – der betroffenen Firmen die Vorfälle bei der Polizei oder der Staatsanwaltschaft. Viel häufiger wurden firmeninterne Massnahmen ergriffen, oft sekundiert durch externe Unterstützung. Dass selbst Flaggschiffe der hiesigen Wirtschaft nicht vor Cyberangriffen gefeit sind, zeigten im Jahr 2023 die Attacken auf die Neue Zürcher Zeitung und die SBB.

Da viele Firmen sich bedeckt halten und Cyberangriffe nicht melden, ist es nicht möglich, den wirtschaftlichen Schaden zu beziffern, den Deepfakes anrichten, zumal nur ein Teil der Online-Attacken auf deren Konto geht. Hingegen existieren Zahlen zum illegalen Markt, der rund um gefälschte Videos entstanden ist: Für ein einfaches Fake-Video musste man 2023 im Darkweb mit Kosten von zwanzig US-Dollar pro Minute rechnen. Neben Dienstleistungen erhält man auf den entsprechenden Plattformen auch Anleitungen, Diskussionsbeiträge und andere Hilfestellungen rund um Deepfakes. Hilfreiche Tipps für den Vertrieb und Kauf von Produkten und Diensten rund um die manipulierten Videos sind vor allem in englisch- und russischsprachigen Darkweb-Foren zu finden, aber auch türkisch-, spanisch- und chinesischsprachige Darknet-Sites verzeichnen diesbezüglich rege Aktivitäten.

Mit welchen Massnahmen die Risiken durch Deepfakes eingedämmt werden könnten, zeigt das abschliessende Kapitel mit den Empfehlungen aus der Studie von TA-SWISS.

Deepfakes im Auge des Gesetzes

In der Schweiz deckt das bestehende Gesetz das meiste Unrecht ab, das mit Deepfakes begangen werden kann. Die Durchsetzung des Rechts steht aber vor erheblichen Schwierigkeiten. Es braucht dazu die internationale Zusammenarbeit.

Kunstfreiheit sowie Meinungs- und Informationsfreiheit gehören zu den Grundfreiheiten und geniessen in der Schweiz einen hohen Stellenwert. Gewährleistet sind sie sowohl durch die Bundesverfassung der Schweizerischen Eidgenossenschaft als auch durch die Europäische Konvention zum Schutze der Menschenrechte und Grundfreiheiten. Auch Deepfakes sind grundrechtlich geschützt. Allerdings darf dieser Schutz eingeschränkt werden, wenn mit den gefakten Inhalten die Rechte anderer geschädigt werden.

Urheberschutz für kreative Leistungen

Synthetische Medien geniessen mithin urheberrechtlichen Schutz, sofern sie als Werk gelten: Darunter fallen dem Urheberrechtsgesetz zufolge auch fotografische, filmische und andere visuelle oder audiovisuelle Produktionen. Der künstlerische oder ästhetische Gehalt des Videos oder Bildes spielt dabei keine Rolle; entscheidend ist der kreative Beitrag eines Menschen. Bei einem vollautomatisiert durch eine KI erzeugten Video fehlt allerdings die schöpferische Leistung. Daher ist es fraglich, ob ohne menschliches Zutun entstandene Videos als Werke zu betrachten und vom Urheberrecht oder von der Kunstfreiheit geschützt sind.

Grenzen der Informationsfreiheit

Auch die Informationsfreiheit hat Grenzen. Bewusst falsche Angaben fallen nicht in den Schutzbereich. Doch ausser im Radio- und Fernsehgesetz, das den Rundfunk zu sachgerechter Information verpflichtet, gibt es kein Gesetz, das den Wahrheitsgehalt von Videos vorschreibt.

Würden indes bedrohliche Deepfakes die Bevölkerung terrorisieren, etwa indem sie diese mit einer bevorstehenden Katastrophe einschüchterten, könnten die Urheber eines solchen Films strafrechtlich belangt werden. Artikel 258 des Schweizer Strafgesetzbuches sieht nämlich Strafen vor für Personen, welche die Bevölkerung durch Androhen oder Vorspiegeln einer Gefahr für Leib, Leben oder Eigentum in Schrecken versetzen.

Identitätsklau, Rufschädigung und Betrug durch Deepfakes

Das Zivilgesetzbuch schützt verschiedene Persönlichkeitsrechte – darunter das Recht am eigenen Bild, an der eigenen Stimme und am eigenen Namen. Deepfakes, die sich der Fotos und Tonaufnahmen einer Person bedienen, ohne dass diese eingewilligt hat, verletzen mithin deren Persönlichkeitsrechte.

Wer in einem Deepfake unvorteilhaft in Szene gesetzt wird, kann sich wehren: Gegen üble Nachrede und Ehrverletzung greifen verschiedene Artikel des Schweizer Strafgesetzbuchs. Auch eine Person, die arglistig getäuscht wird – etwa, um sie zu einer Geldüberweisung oder zur Preisgabe vertraulicher Informationen zu veranlassen, kann sich auf das Strafgesetzbuch berufen.

Das Strafrecht ahndet die Verbreitung pornographischer Inhalte – ein Tatbestand, den ein grosser Teil der produzierten Deepfakes erfüllen dürfte. Es ist strafbar, pornographische Ton- und Bildaufnahmen öffentlich auszustellen oder jemandem unaufgefordert anzubieten. Wird ein Deepfake einer einzelnen Person aufgedrängt, kann auch der Straftatbestand der sexuellen Belästigung erfüllt sein. Gegen das Phänomen des Rachepornos richtet sich ein neuerer Artikel im Strafgesetzbuch, der das unbefugte Weiterleiten von nicht öffentlichen sexuellen Inhalten unter Strafe stellt.

Ausgeklügelte Urkundenfälschung

Nicht selten kommen Aufzeichnungen von Überwachungskameras oder Bodycams in Gerichtsverfahren als Beweismittel zum Einsatz. Mittels Deepfake-Technologien könnten solche Videos verfälscht werden – oder es könnten umgekehrt Deepfakes für falsche Alibis verwendet werden. In Form von manipulierten Dokumenten sind Urkundenfälschungen nichts Neues; mit synthetisch generierten Videos und Audioaufzeichnungen wird ihnen aber ein weiteres Kapitel hinzugefügt.

Angesichts von Deepfakes wird die Bewertung audiovisueller Beweismittel noch anspruchsvoller. Doch mit welchen Mitteln auch immer prozessrechtliche Akten, Videos oder Tonaufzeichnungen manipuliert werden: Wer eine Urkunde fälscht, begeht eine Straftat. Und wer wider besseres Wissen Anschuldigungen gegen eine Person erhebt, kann seinerseits aufgrund falscher Anschuldigungen bzw. Irreführung der Rechtspflege zur Rechenschaft gezogen werden.

Synthetische Medien als Hilfsmittel bei der Strafverfolgung

Unter Fachleuten des Rechts wird diskutiert, synthetische Medien als Instrument gegen Kriminelle einzusetzen. So muss in verdeckten Ermittlungen oft kinderpornografisches Material hochgeladen werden, damit die entsprechenden Online-Foren infiltriert werden können. Allerdings dürfen Ermittler selbst dann keine Delikte begehen, wenn sie Kriminelle jagen – und hierzulande stellt das Verbreiten «fiktiver» Kinderpornographie eine Straftat dar. Zudem ist sogar das Erzeugen eines gänzlich synthetischen kinderpornografischen Deepfakes heikel, weil die Herstellung eines realistisch wirkenden Videos Trainingsdaten benötigt, die aus Bildern von tatsächlich begangenem Kindesmissbrauch bestehen. Somit darf die Polizei bei ihren Ermittlungen keine kinderpornographischen Deepfakes verwenden.

Schliesslich wäre es denkbar, Deepfakes in der Strafverfolgung einzusetzen – etwa, um anhand von Handyvideos, Daten aus Überwachungskameras und Körperscans den Hergang eines Verbrechens zu rekonstruieren. Ein solches Verfahren wirft aber eine Reihe rechtlicher Frage auf. Problematisch ist, dass eine solche Rekonstruktion zwar objektiv wirkt, dabei aber allein auf den Annahmen der Strafverfolgung beruht. Unklar ist zudem, wie den Beschuldigten ihr Recht gewährt werden kann, an sämtlichen Verfahrensschritten – inklusive der «virtuellen Tatortbegehung» – teilzunehmen, und wie schliesslich die digitalen Beweise zu den Akten gelegt werden sollen.

Internationale Zusammenarbeit im Kampf gegen globalisierte Taten

Die Schweizer Gesetzgebung deckt somit die meisten Delikte ab, die mit Deepfakes begangen werden können.

Die Durchsetzung von Schweizer Recht dürfte allerdings oft an hohen Hürden scheitern. Denn die Urheberschaft von Deepfakes ist meistens schwer zu ermitteln. Und selbst wenn es gelingt, die Täterschaft ausfindig zu machen, nützt das wenig, wenn sich ein Deepfake schon tausendfach verbreitet und in alle Winde verstreut hat. Die meisten Deepfakes sind ausländischen Ursprungs und werden auf Plattformen ausserhalb der Schweiz hochgeladen. Zudem ziehen Straftaten, die im Internet begangen werden, meistens aufwendige Verfahren nach sich. Häufig sind mehrere Personen beteiligt, die es zu identifizieren gilt, hinzu kommen unklare Zuständigkeiten und überlastete Behörden der Strafverfolgung.

Sowohl von Rechtshilfeabkommen als auch von einer vertieften internationalen Kooperation beim Datenaustausch versprechen sich Fachleute Verbesserungen bei der grenzüberschreitenden Rechtsdurchsetzung. In der Europäischen Union soll die Verordnung namens «Gesetz über Digitale Dienste» (Digital Services Act DSA) für den besseren Schutz der Nutzerinnen und Nutzer im Internet sorgen. Unter anderem werden die Plattformen verpflichtet, illegale Inhalte zu bekämpfen. Ausserdem müssen sie es den Userinnen und Usern ermöglichen, Inhalte zu melden, und sie haben mit sogenannten Trusted Flaggers zu kooperieren. Bei diesen 'vertrauenswürdigen Hinweisgebern' handelt es sich um Institutionen, die rechtswidrige Inhalte aufspüren und der Plattform berichten sollen. Jüngst hat die EU auch einen Rechtsakt über künstliche Intelligenz verabschiedet, der für Deepfakes eine Transparenzpflicht vorsieht.

Online-Plattformen wie soziale Netzwerke selber sind nicht untätig geblieben. Mehrere von ihnen haben Community-Richtlinien ausgearbeitet, die digitale Fälschungen und irreführende Informationen verbieten. Zudem haben 34 grosse Firmen – darunter Meta, Google, Microsoft und TikTok – einen Verhaltenskodex unterschrieben, mit dem sie sich zur Bekämpfung von Fehlinformationen verpflichten. Allein auf die Selbstregulierung der grossen Plattformen zu bauen, würde indes dem öffentlichen Interesse kaum gerecht. Denn letztlich fehlt es bei der Festlegung von Löschungskriterien grundsätzlich an demokratischer Mitwirkung und Transparenz. Die Gefahr einseitiger Machtausübung ist nicht von der Hand zu weisen.

Realitätsverzerrungen korrigieren: Einige Empfehlungen für den Umgang mit Deepfakes

Allein durch regulatorische oder durch technische Einzelmassnahmen lassen sich unliebsame Folgen von Deepfakes nicht verhindern oder auch nur eindämmen. Vielmehr braucht es einen Mix aus Vorkehrungen auf unterschiedlichen Ebenen und viel Selbstverantwortung, um auch vom Potential der synthetischen Medien profitieren zu können.

Da die meisten manipulierten Videos ihr Publikum über die grossen Online-Plattformen finden, kommt diesen bei der Regulierung von Deepfakes eine Schlüsselrolle zu. Gefordert sind zudem die Behörden, die Kommunikationsbranche, die Bildung – und nicht zuletzt alle Bürgerinnen und Bürger.

Selbstverantwortung wahrnehmen

In sämtlichen Branchen sollte die Aus- und Weiterbildung zu Medien- und Informationskompetenz ganz oben auf der Prioritätenliste stehen. Bürgerinnen und Bürger wiederum sollten eigenverantwortlich die Bildungs- und Aufklärungsangebote verschiedener Stellen nutzen. Selbstverantwortung ist auch bei der Bewertung, Weiterverbreitung und nicht zuletzt bei der Herstellung von Deepfakes ein Gebot der Stunde. Zudem sollte jeder und jedem bewusst sein, dass das Hochladen von Bildern und Sprachaufnahmen die Produktion von Deepfakes begünstigen kann. Der Grundsatz, wonach das Internet nicht vergisst, gilt besonders mit Blick auf Deepfakes.

Wer sich gerne im Internet Videos anschaut oder ein solches über soziale Medien zugespielt erhält, sollte stets in Betracht ziehen, dass es sich um eine Fälschung handeln kann. Skepsis ist insbesondere dann angezeigt, wenn die Aufzeichnung emotional aufwühlt oder besonders spektakulär ist.

Plattformen in die Pflicht nehmen und Opferschutz stärken

Der Staat sollte sich dafür starkmachen, Plattformen zur Löschung oder Sperrung gemeldeter Deepfakes zu verpflichten. Zudem sollten Plattformbetreiber dazu angehalten werden, ein Meldesystem für Deepfakes einzurichten. Auch Transparenzvorgaben und Widerspruchsmöglichkeiten würden die Rechte sowohl der Opfer von Deepfakes als auch der Betroffenen ungerechtfertigter Löschungen stärken. Um solche Massnahmen durchzusetzen, ist die internationale Zusammenarbeit unabdingbar; mithin müssten zusätzliche Kooperationsinstrumente mit anderen Staaten geschaffen werden. Zudem sollte sich die Schweiz dafür einsetzen, dass international gültige Normen und Regeln gegen problematische Deepfakes und gegen Cyberkriminalität definiert werden.

Im Konflikt mit grossen Online-Plattformen haben Einzelpersonen in der Regel das Nachsehen. Daher braucht es spezialisierte Fachstellen, welche die Opfer von Deepfakes – oder die Betroffenen ungerechtfertigter Löschungen – beraten und unterstützen. Opferberatungsstellen für Cyberdelikte sollten von Bund und Kantonen mit genügenden personellen und finanziellen Mitteln ausgestattet werden. Auch sollte die Schweiz Trusted Flaggers staatlich anerkennen, sodass deren Meldungen von Deepfakes im Internet Vorrang eingeräumt werden müsste; allenfalls wäre eine finanzielle Unterstützung solcher Hinweisgeber zu erwägen.

Den technischen Fortschritt für die Verteidigung nutzen

Eine breite Debatte über Authentifizierungs- und Kennzeichnungsverfahren ist zu begrüssen. Fortschrittliche Methoden, insbesondere die Mehr-Faktor-Authentifizierung, können dazu beitragen, Täuschungsversuche mittels Stimm- oder Gesichts-Deepfake zu vereiteln. Dabei ist es wichtig, nach Möglichkeit auf die am meisten ausgereiften Authentifizierungsverfahren zurückzugreifen. Denn Cyberkriminelle arbeiten ihrerseits daran, die Schutzmassnahmen zu überwinden.

Angesichts der rasanten Entwicklung der Deepfake-Techniken gilt es, alle denkbaren Instrumente zu nutzen, um Missbrauch zu verhindern. Selbst Hilfsmittel, die derzeit noch wenig wirksam sind wie die Deepfake-Detektoren, können ein Element im Mosaik eines umfassenden Schutzes darstellen. Zudem empfiehlt sich eine möglichst robuste Gestaltung bestehender Sicherheitsmassnahmen.

Aufklärung über die Risiken – und über die Nutzen

Zurzeit haben erst wenige Leute Erfahrungen mit Deepfakes gesammelt, und viele wissen kaum darüber Bescheid. Informationen an Schulen und in den Medien sollten das Bewusstsein für das Phänomen schärfen; hilfreich wären insbesondere Tipps, wie Quellen verifiziert und die Plausibilität von Videos hinterfragt werden können. Schulen sollten prüfen, ob eine Auseinandersetzung mit Deepfakes unter die Ziele des Lehrplans 21 zur Stärkung der Medienkompetenz fallen könnte.

Trotz der Risiken, über die es aufzuklären gilt, sollten die Potenziale synthetischer Videos nicht erstickt werden. Daher ist auf eine sorgfältige Wortwahl zu achten. Denn die Menschen bringen den Ausdruck «Deepfakes» viel weniger mit Chancen in Verbindung als neutralere Begriffe wie «synthetische Medien».

Mitglieder der Begleitgruppe

- **Prof. Dr. Reinhard Riedl**, Berner Fachhochschule BFH, Präsident der Begleitgruppe, Mitglied des Leitungsausschusses von TA-SWISS
- **Dr. Bruno Baeriswyl**, Datenschutzexperte, Präsident des Leitungsausschusses von TA-SWISS
- **Cornelia Diethelm**, Centre for Digital Responsibility
- **Prof. Dr. Rainer Greifeneder**, Leiter der Abteilung Sozialpsychologie, Universität Basel
- **Thomas Häussler**, Abteilung Medien / Sektion Grundlagen Medien, Bundesamt für Kommunikation BAKOM
- **Andrea Hauser**, Informatikerin, Sicherheitsexpertin Cybersecurity, Sicherheitsfirma Scip
- **Erich Herzog**, Rechtsanwalt, Mitglied der Economiesuisse Geschäftsleitung
- **Prof. Dr. Selina Ingold**, IDEE Institut für Innovation, Design & Engineering, Ostschweizer Fachhochschule
- **Melanie Kömle Bender**, Mediendokumentalistin, SRF Schweizer Radio und Fernsehen
- **Thomas Müller**, Wissenschaftsjournalist, Mitglied des Leitungsausschusses von TA-SWISS
- **Prof. Dr. René Schumann**, HES-SO Valais-Wallis, Forschungsinstitut Informatik
- **Prof. Dr. Giatgen Spinas**, Universität Zürich, Mitglied des Leitungsausschusses von TA-SWISS
- **Dr. Stefan Vannoni**, Ökonom, CEO cemsuisse, Mitglied des Leitungsausschusses von TA-SWISS

Projektmanagement TA-SWISS

- **Dr. rer. soc. Elisabeth Ehrensperger**, Geschäftsführung
- **Dr. Laetitia Ramelet**, Projektleitung
- **Dr. Lucienne Rey**, Projektleitung
- **Fabian Schluep**, Kommunikation

Impressum

Augen und Ohren auf dem Prüfstand
Kurzfassung der Studie von TA-SWISS «Deepfakes und manipulierte Realitäten»
TA-SWISS, Bern 2024
TA 81A/2024

Autorin: Lucienne Rey
Produktion: Laetitia Ramelet und Fabian Schluep
Gestaltung und Illustrationen: Hannes Saxer, Bern
Druck: Jordi AG – Das Medienhaus, Belp

TA-SWISS – Stiftung für Technologiefolgen-Abschätzung

Neue Technologien bieten oftmals entscheidende Verbesserungen für die Lebensqualität. Zugleich bergen sie mitunter aber auch neuartige Risiken, deren Folgen sich nicht immer von vornherein absehen lassen. Die Stiftung für Technologiefolgen-Abschätzung TA-SWISS untersucht die Chancen und Risiken neuer technologischer Entwicklungen in den Bereichen «Biotechnologie und Medizin», «Digitalisierung und Gesellschaft» sowie «Energie und Umwelt». Ihre Studien richten sich sowohl an die Entscheidungstragenden in Politik und Wirtschaft als auch an die breite Öffentlichkeit. Ausserdem fördert TA-SWISS den Informations- und Meinungsaustausch zwischen Fachleuten aus Wissenschaft, Wirtschaft, Politik und der breiten Bevölkerung durch Mitwirkungsverfahren. Die Studien von TA-SWISS sollen möglichst sachliche, unabhängige und breit abgestützte Informationen zu den Chancen und Risiken neuer Technologien vermitteln. Deshalb werden sie in Absprache mit themenspezifisch zusammengesetzten Expertengruppen erarbeitet. Durch die Fachkompetenz ihrer Mitglieder decken diese Begleitgruppen eine breite Palette von Aspekten der untersuchten Thematik ab.

Die Stiftung TA-SWISS ist ein Kompetenzzentrum der Akademien der Wissenschaften Schweiz.

TA-SWISS
Stiftung für Technologiefolgen-Abschätzung
Brunngasse 36
CH-3011 Bern
info@ta-swiss.ch
www.ta-swiss.ch

mitglied der
akademien der
wissenschaften schweiz