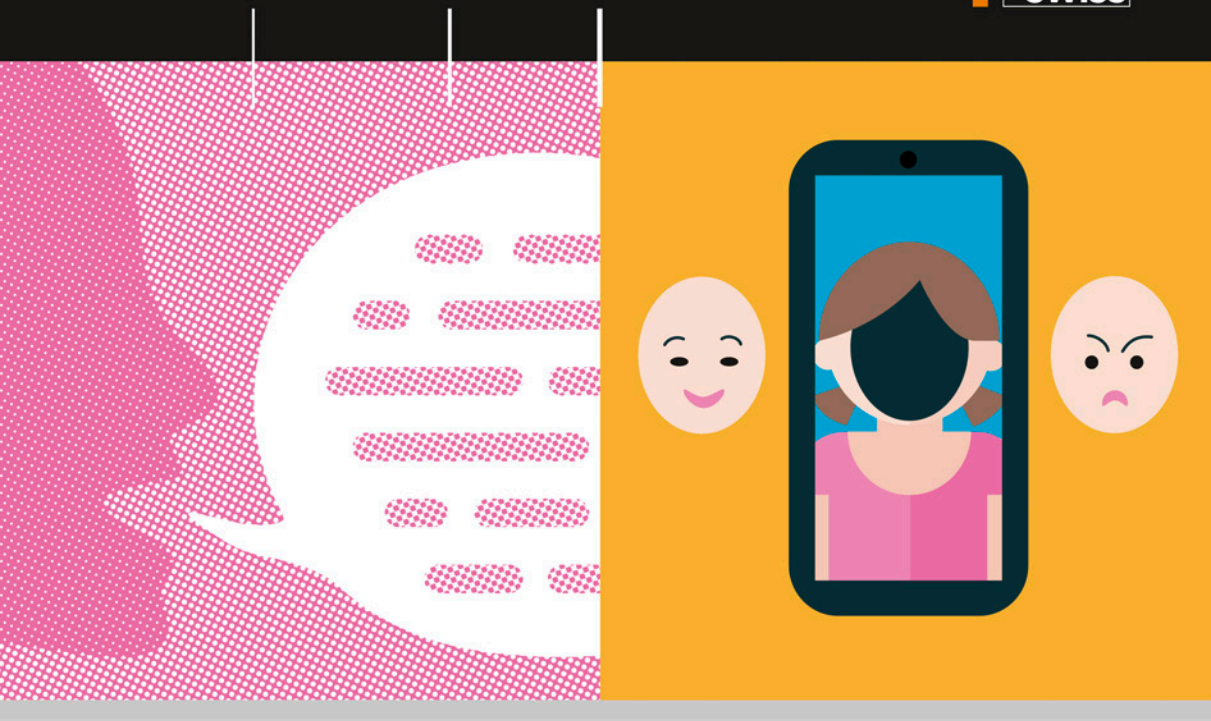

*Murat Karaboga, Nula Frei, Manuel Puppis, Daniel Vogler, Patric Raemy, Frank Ebbers, Greta Runge, Adrian Rauchfleisch, Gabriele de Seta, Gwendolyn Gurr, Michael Friedewald, Sophia Rovelli*

# Deepfakes und manipulierte Realitäten

## Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz

v/d/f

TA-SWISS 81/2024

Brunngasse 36
CH-3011 Bern
www.ta-swiss.ch

*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*

# Deepfakes und manipulierte Realitäten

## Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz

**Bibliografische Information der Deutschen Nationalbibliothek**
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.



**Zitiervorschlag**
Karaboga, M., Frei, N., Puppis, M., Vogler, D., Raemy, P., Ebbers, F., Runge, G., Rauchfleisch, A., de Seta, G., Gurr, G., Friedewald, M. & Rovelli, S. (2024):
Deepfakes und manipulierte Realitäten. Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz.
TA-SWISS Publikationsreihe (Hrsg.): TA 81/2024. Zolllikon: vdf.

**Zitiervorschlag einzelner Buchbeitrag (Beispiel)**
Vogler, D., Rauchfleisch, A. & de Seta, G. (2024): Wahrnehmung von Deepfakes in der Schweizer Bevölkerung.
In: Karaboga, M., Frei, N., Puppis, M., Vogler, D., Raemy, P., Ebbers, F., Runge, G., Rauchfleisch, A., de Seta, G., Gurr, G., Friedewald, M. & Rovelli, S.:
Deepfakes und manipulierte Realitäten. Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz.
TA-SWISS Publikationsreihe (Hrsg.): TA 81/2024. Zolllikon: vdf. S. 125–151.

Coverabbildungen:
© Rechts: Hannes Saxer, Bern



ISBN 978-3-7281-4185-9 (Printausgabe)

Download open access:
ISBN 978-3-7281-4186-6 / DOI 10.3218/4186-6

www.vdf.ch
verlag@vdf.ch

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Zusammenfassung

Ein Deepfake ist ein mithilfe von KI-Techniken synthetisierter oder manipulierter Audio- oder (Bewegt-)Bildinhalt, der authentisch erscheint und in dem zumeist eine Person etwas sagt oder tut, was sie nie gesagt oder getan hat. Seit der ersten Bezeichnung synthetischer und manipulierter Medieninhalte als «Deepfake» im Jahr 2017 hat der Begriff einen festen Platz in der politischen und medialen Debatte eingenommen.

Mehrere Jahre nach ihrem ersten Auftreten lässt sich inzwischen eine gemischte Bilanz ziehen: Mehrere politische Deepfakes und Deepfake-basierte Betrugsfälle, wie zuletzt Schockanrufe, scheinen manche Befürchtungen zu bestätigen. Auf der anderen Seite sind auch eine Reihe von nützlichen Anwendungen entstanden, die auf synthetischen und manipulierten Medieninhalten basieren. Die befürchtete massenhafte Nutzung von Deepfakes in Desinformationskampagnen oder gar die grosse Informationsapokalypse sind aber bislang ausgeblieben.

Unsere Studie verfolgt das Ziel, die Chancen und Risiken von Deepfakes für die Schweiz abzuschätzen. Angesichts der Fortschritte bei Deepfake-Technologien stellt sich zunächst die Frage, wie der *aktuelle technologische Stand* ist. Bislang kaum erforscht ist, *was die Bürgerinnen und Bürger über Deepfakes wissen* und wie sie die Chancen und Risiken einschätzen. Des Weiteren wird untersucht, wie die Nutzung von Deepfakes *rechtlich* zu bewerten ist: Was ist erlaubt und wo sind die Grenzen? Und reicht die bestehende Gesetzgebung z.B. aus, um die Bevölkerung vor der missbräuchlichen Verwendung von Deepfakes zu schützen? Als ein wichtiger Baustein im Umgang mit Deepfakes wird immer wieder das verantwortungsbewusste Handeln des *Journalismus* diskutiert. Doch wie werden Deepfakes im Schweizer Journalismus wahrgenommen und wie gehen Journalistinnen und Journalisten mit Deepfakes um? Ebenfalls viel diskutiert ist die Frage der möglichen *politischen Implikationen* von Deepfakes. Hier fragen wir uns konkret, welche Rolle Deepfakes in der Politik spielen könnten und welche Handlungsoptionen sich der Politik und anderen Akteuren bieten, um unerwünschte Auswirkungen zu vermeiden. Besonders viel Potenzial wird jüngst der Verwendung von Deepfakes für *wirtschaftliche Zwecke* zugerechnet: In diesem Zusammenhang untersuchen wir daher sowohl, welche wirtschaftlichen Chancen Deepfake-Technologien mit sich bringen können, als auch, für welche missbräuchlichen wirtschaftlichen Zwecke sie eingesetzt werden können und welche

präventiven und reaktiven Schutzmassnahmen (potenziell) Betroffene treffen können. Schliesslich verfolgt unsere Studie nicht nur das Ziel, Orientierungswissen über die Chancen und Risiken von Deepfake-Technologien zu bieten, sondern auch *Handlungsmöglichkeiten aufzuzeigen*, wie ein verantwortungsbewusster Umgang mit Deepfakes aussehen kann, sodass die Wahrnehmung der Chancen erhalten bleibt, während Risiken, wenn nicht vermieden, so doch zumindest reduziert werden. Den hier aufgeworfenen Fragen widmet sich die Studie in einzelnen vertiefenden Kapiteln, deren Kernergebnisse im Folgenden kurz vorgestellt werden:

- Ist- und Trendanalyse zu Deepfake-Technologien
- Wahrnehmung von Deepfakes in der Bevölkerung
- Deepfakes im Recht
- Deepfakes im Journalismus
- Deepfakes in der Politik
- Deepfakes in der Wirtschaft
- Empfehlungen

## Ist- und Trendanalyse zu Deepfake-Technologien

Die Untersuchung der aktuellen und absehbaren technischen Möglichkeiten zur Erstellung von Deepfakes zeigt, dass in den vergangenen Jahren beachtliche Fortschritte bei der KI-basierten Synthetisierung und Manipulation von Bild-, Audio- und Textinhalten erzielt wurden.

Im Bereich *bildbasierter* Deepfakes stehen gegenwärtig fünf Techniken zur Verfügung:

- *Facial reenactment*: Die Manipulation des Gesichtsausdrucks,
- *Face morphing*: Die Verschmelzung von mehreren Gesichtern,
- *Face swapping*: Das Tauschen oder Ersetzen eines Gesichts durch ein anderes,
- *Gesichtsgeneration*: Die Erschaffung von Gesichtern, die in der Realität nicht existieren, und

- *Full body puppetry*: Das sog. Ganzkörperpuppenspiel, bei dem die Pose oder Bewegungen eines Körperteils oder des gesamten Körpers verändert werden.

Im Bereich von Deepfake-*Audios* hat sich die Entwicklung in den vergangenen Jahren in Richtung zunehmend glaubwürdiger Resultate bewegt, die mit zunehmend weniger Trainingsmaterial erzielt werden können.

So ist bereits heute die Erschaffung einer breiten Palette an Deepfake-Inhalten möglich. Moderne Deepfake-Technologien erlauben das Kopieren der Sprechweise, des Gesichtsausdrucks und sogar der Körperbewegungen eines Menschen – immer vorausgesetzt, die Erstellenden verfügen über ausreichend Inputdaten (z.B. Sprachsamples oder Gesichtsfotos) und Computerleistung. Mit modernen Textgeneratoren ist zudem die Imitation der Schreibweise und Sprache eines Menschen möglich, was als Input für Deepfakes genutzt werden kann. Auch können Hintergründe in immer besserer Qualität vollständig synthetisiert und verändert werden. Damit ist nicht nur die Erschaffung von Deepfakes möglich, die Menschen abbilden, sondern auch realistische Szenarien wie einen gefälschten Flugzeugabsturz oder eine Naturkatastrophe. Deepfake-Audios mit hoher Qualität lassen sich schon heute mit vergleichsweise geringem Aufwand produzieren. Ebenso können mit KI-Bildgeneratoren täuschend echte Deepfake-Bilder von Personen produziert werden, die dem Generator bekannt sind, also insb. von Personen des öffentlichen Lebens. Durch das Trainieren der Generatoren mit weiterem Material oder durch das Hinzufügen eigener Bilder zu einem Textbefehl ist auch die Generierung von Abbildungen beliebiger weiterer Personen möglich.

Darüber hinaus bestätigen von uns selbst durchgeführte Versuche zur Erstellung von Deepfake-Videos, dass trotz enormer Fortschritte bei der Technologie auch weiterhin ein grosses technisches Know-how und enorme Rechenkapazitäten bzw. Zeit und Geld erforderlich sind, um glaubwürdige Deepfake-Videos zu erschaffen. Allerdings werden Deepfake-Generatoren zunehmend einfacher bedienbar und grundlegend ändern könnte sich die Situation, wenn Text-zu-Video-Verfahren ausgereift und für breite Massen der Bevölkerung verfügbar sind. Mit diesen wäre es möglich, beliebige Deepfake-Videos unter Eingabe von Textbefehlen zu erstellen, ähnlich wie mit aktuellen KI-basierten Text- oder Bildgeneratoren wie ChatGPT oder Midjourney. Es kann davon ausgegangen werden, dass die bekanntesten Programme dieser Art Schutzvorkehrungen treffen werden, um schädliche und unethische Nutzungen zu verhindern. Andererseits können solche Schutzvorkehrungen umgangen werden. Die Entstehung weiterer Generatoren, die auf ethische Leitlinien verzichten, ist naheliegend. Auch

wenn also keine akute Gefahr besteht, dass die (sozialen) Medien mit glaubwürdigen Deepfake-Videos «überflutet» werden, wird deren Erstellung künftig deutlich einfacher werden. Vor dem Hintergrund dieser Ergebnisse gehen wir davon aus, dass Deepfake-Videos in der nahen Zukunft vor allem durch Einzelne oder organisierte Akteure produziert werden, die über das erforderliche Know-how und die technischen Rechenkapazitäten verfügen.

Die folgenden Kapitel werden allerdings zeigen, dass keine grosse Anzahl von Deepfakes für schädliche Auswirkungen notwendig ist. Ein einzelner Deepfake kann bereits grossen Schaden politischer, rechtlicher, sozialer oder wirtschaftlicher Art anrichten. Und wie z.B. die Reaktionen auf ein manipuliertes Foto von Papst Franziskus in Daunenjacke im Jahr 2023 demonstriert haben, ist nicht einmal eine Irreführungs- oder Schädigungsabsicht notwendig. Menschen können auch von Medieninhalten in die Irre geführt werden, die zu Unterhaltungszwecken erstellt wurden.

Technische Massnahmen, die als Abhilfe diskutiert werden, umfassen Methoden zur Authentifizierung von originalen Inhalten, zur Kennzeichnung von Deepfake-Inhalten sowie zur Erkennung von Deepfake-Videos. Alle diese Massnahmen haben Schwächen. Gegen die Authentifizierung vertrauenswürdiger Inhalte und die Kennzeichnung von Deepfake-Inhalten spricht ihre prinzipielle technische Überwindbarkeit. Fälschlich als vertrauenswürdig signierte Deepfakes könnten nicht nur selbst grossen Schaden anrichten, sondern auch das Vertrauen in signierte Dokumente grundsätzlich beschädigen. Die Überwindung technischer Schutzvorkehrungen könnte vermutlich vor allem besonders ressourcenstarken (staatlichen) Akteuren gelingen. Dies könnte jedoch zugleich bedeuten, dass die Masse der kursierenden Deepfakes gekennzeichnet und die Masse der vertrauenswürdigen Inhalte immerhin als solche signiert und erkennbar wäre. In Kombination mit weiteren Massnahmen könnte das Signieren trotz seiner Schwächen somit ein Element einer Gesamtstrategie zur Eindämmung unerwünschter Wirkungen von Deepfakes sein.

Unsere Auswertung des Stands der Literatur und eigene Tests zeigen, dass frei verfügbare Deepfake-Detektoren derzeit nicht zuverlässig sind. Dafür, dass Verbesserungen nicht zu erwarten sind, sprechen Einschätzungen, wonach die Generatoren den Detektoren stets einen Schritt voraus sein werden. Angesichts der Investitionen in die Entwicklung von Detektoren und der technischen Herausforderungen, denen *GAN*s (Generative Adversarial Networks) und andere Deepfake-Technologien gegenüberstehen, besteht allerdings die Möglichkeit, dass die Genauigkeit künftiger Verfahren dennoch zunimmt.

## Wahrnehmung von Deepfakes in der Schweizer Bevölkerung

Ein bislang kaum erforschtes Thema ist die Einstellung der Bevölkerung zu Deepfakes. Mit einer repräsentativen Befragung haben wir untersucht, wie Deepfakes in der Schweizer Bevölkerung wahrgenommen werden. Dazu haben wir erhoben, welche Erfahrungen Schweizerinnen und Schweizer mit Deepfakes haben, wie sie die Chancen und Risiken von Deepfake-Technologien einschätzen und ob sie in der Lage sind, Deepfakes von realen Videos zu unterscheiden.

Deepfakes sind im September 2023 für viele Menschen noch ein unbekannter Begriff. Lediglich etwas mehr als die Hälfte der Befragten hat schon von Deepfakes gehört. Knapp die Hälfte von allen Befragten hat nach eigenen Angaben bereits Deepfakes gesehen. Nur eine sehr kleine Minderheit hat schon Erfahrungen mit dem Erstellen (2 %) und Verbreiten (3 %) von Deepfakes. Das bedeutet einerseits, dass die Wahrnehmung des Themas in erster Linie durch die Medienberichterstattung geprägt ist und nicht auf eigenen Erfahrungen beruht. Dies eröffnet aber auch die Möglichkeit, mit Bildungs- und Informationsangeboten zur Thematik einen sinnvollen Umgang mit Deepfake-Technologie zu vermitteln.

Deepfakes werden von der Schweizer Bevölkerung stark mit Risiken assoziiert, wobei es kaum Unterschiede zwischen soziodemografischen Gruppen gibt. Allerdings spielt der vermutete Effekt von Deepfakes auf die Meinung von Drittpersonen eine Rolle. Je eher man davon ausgeht, dass andere Menschen von Deepfakes beeinflusst werden, desto höher schätzt man selber das Risiko von Deepfakes für die Gesellschaft ein. Es scheint also eine eher diffuse Vorstellung über die Effekte von Deepfakes auf die Gesellschaft zu bestehen, welche die Risikowahrnehmung erhöhen. Die Verwendung der Bezeichnung «synthetische Medien» statt des Begriffs «Deepfakes» wirkt sich hingegen positiv auf die Wahrnehmung von Chancen von Deepfake-Technologien aus. In der Summe deuten die Befunde darauf hin, dass noch viel Unsicherheit beim Thema Deepfakes in der Bevölkerung besteht. Ein relevanter Befund ist zudem, dass Frauen individuelle Risiken – also die Gefährdung der Privatsphäre oder das Risiko, selbst Opfer eines Deepfake zu werden – höher einschätzen als Männer.

Unsere Experimente zeigen, dass Menschen in der Schweiz Deepfakes kaum von realen Videos unterscheiden können, insbesondere, wenn diese von guter bis sehr guter Qualität sind. Umso wichtiger sind übergeordnete Medienkompetenzen. Unsere Studie zeigt, dass *Social Media Literacy*, also Kompetenz im Umgang mit neuen Medien im Internet, einen positiven Effekt auf die Deepfake-Erkennungskompetenz hat. Die generelle Medienkompetenz und Internet-

Skills haben hingegen keinen Effekt. Unsere kurze Literacy-Intervention, also eine Hilfestellung zum Erkennen von Deepfakes unmittelbar vor dem Zeigen der Videos, hatte keinen Effekt auf die Erkennungskompetenz. Allerdings tritt auch kein *Backfire-Effekt* auf. Die Literacy-Intervention führt also nicht dazu, dass die Befragten überkritisch wurden und in der Folge auch reale Videos eher für Deepfakes hielten.

## Deepfakes im Recht

Das Recht hat eine doppelte Rolle in Bezug auf Deepfakes: Einerseits schützt es Deepfakes als künstlerischen Ausdruck oder als Meinungsäusserung, andererseits kommt ihm aber auch die wichtige Rolle zu, Schutz vor schädlichen und/oder unerwünschten Auswirkungen von Deepfakes zu bieten, insbesondere durch die Mittel des Strafrechts, des Urheberrechtsschutzes und des Persönlichkeits- und Datenschutzes, aber auch durch öffentlich-rechtliche (Medien-)Regulierung. So können Deepfakes resp. deren Urheber auf der einen Seite einen Schutz der Meinungs-, Informations- und Kunstfreiheit beanspruchen, da selbst unwahre, falsche oder irreführende Aussagen zu einem gewissen Grad grundrechtlichen Schutz geniessen. Zudem können Deepfakes unter Umständen urheberrechtlich geschützt sein. Dieser Schutz ist aber nicht absolut: Grundrechte können unter bestimmten Voraussetzungen eingeschränkt werden, namentlich zum Schutz der Rechte Dritter. Daher liegt ein Schwerpunkt der rechtlichen Analyse in dieser Studie auf dem rechtlichen Schutz vor den schädlichen Anwendungen bzw. Auswirkungen von Deepfakes.

In der Schweiz gibt es derzeit keine spezifische Regulierung von Deepfakes, sondern es sind je nach Anwendungsform unterschiedliche allgemeine Rechtsvorschriften zu beachten. «Deepfakes» und andere in diesem Zusammenhang verwendete Bezeichnungen sind denn auch keine Rechtsbegriffe, werden aber teilweise von bestehenden Konzepten erfasst. Zu unterscheiden ist zwischen zivilrechtlichen und strafrechtlichen Vorschriften sowie deren verfahrensrechtlichen Durchsetzung und den öffentlich-rechtlichen Vorgaben.

Viele der «unerwünschten» Anwendungsformen von Deepfakes werden durch das geltende Recht erfasst werden. Im Zivilrecht sind dabei namentlich Persönlichkeitsverletzungen möglich, zudem stellt die Herstellung von Deepfakes normalerweise auch eine Verletzung von urheberrechtlichen und datenschutzrechtlichen Ansprüchen dar. Einige Anwendungen von Deepfakes fallen auch unter das Strafrecht, wobei insbesondere Ehrverletzungsdelikte, Erpressung oder

Betrug sowie Pornografie, sexuelle Belästigung oder Kinderpornografie infrage kommen. Wie der neu geschaffene Straftatbestand des Identitätsmissbrauchs bei Deepfakes angewendet wird, ist derzeit noch offen. Jedoch gibt es Graubereiche, namentlich bei der strafrechtlichen Beurteilung von «peinlichen» oder «freizügigen» Deepfakes als Pornografie oder Belästigung. Es zeigt sich auch, dass gerade bei der Verwendung von Deepfakes als bildbasierte sexualisierte Gewalt der Unrechtsgehalt, der in der Verletzung der Ehre in Kombination mit der sexuellen Selbstbestimmung liegt, nicht vom geltenden Strafrecht erfasst ist.

Die grössten Herausforderungen zeigen sich im Bereich der Durchsetzung zivil- und strafrechtlicher Ansprüche. Im Internet sind die Täter häufig unbekannt oder haben ihren Sitz im Ausland. Auch ein Vorgehen gegen Plattformen ist mit prozessualen Unsicherheiten behaftet. Aufgrund der raschen und potenziell massenhaften Verbreitung von Inhalten im Internet ist es fast unmöglich, einen einmal in Umlauf geratenen Deepfake mit rechtlichen Mitteln «aufzuhalten». Schliesslich können insbesondere im Zivilverfahren die Verfahrenskosten auf Betroffene von Deepfakes abschreckend sein.

Deepfakes können auch im Rechtsalltag eine Rolle spielen, etwa als Beweismittel in Gerichtsverfahren. Das Einbringen von gefälschten Urkunden vor Gericht ist per se keine neue Entwicklung. Programme wie Photoshop ermöglichen bereits seit vielen Jahren die Manipulation von Texten oder Fotos. Deepfake-Technologie ist eine neue technische Variante, die das Fälschen von Video und Ton erleichtert, aber nicht grundsätzlich neue Fragen aufwirft. Die bestehenden Rechtsregeln für den Umgang mit gefälschten Beweismitteln sind daher anwendbar.

Denkbar ist auch, dass die Deepfake-Technologie von den Strafverfolgungsbehörden genutzt wird, um Straftaten aufzuklären. So kann beispielsweise mittels Deepfake-Technologie aus Videos von Handys und Überwachungskameras sowie Körperscans ein virtueller Tatort kreiert oder ein Tathergang rekonstruiert werden. Kontrovers diskutiert wird die Frage, ob verdeckte Ermittler computergenerierte Kinderpornografie verwenden dürfen, um Zutritt zu Onlineforen zu erhalten, die ein Hochladen von eigenen Bildern oder Videos voraussetzen. Während letztere Frage vor allem darum problematisch ist, weil nach aktuellem Recht die Strafverfolgungsbehörden auch zur Aufklärung von Straftaten nicht selber Straftaten begehen dürfen (auch mit Deepfakes erstellte Kinderpornografie fällt u.E. unter die Definition der verbotenen Kinderpornografie), so sind erstere Fragen mit gewissen verfahrensrechtlichen Unsicherheiten behaftet, namentlich der Frage, wie die Gewährung der Teilnahmerechte in allen Verfahrensstadien und somit auch bei der «virtuellen Tatortbegehung» sichergestellt

werden kann, die Frage der Überprüfbarkeit der derart erhobenen Beweise oder die Frage, wie und in welcher Form digital erhobene Beweise künftig zu den Akten gelegt werden.

Schliesslich stellt sich die Frage nach der Regulierung von Deepfakes: Einige Staaten wie etwa die USA verfolgen technologiespezifische Regulierungsansätze, die speziell Deepfakes regeln (etwa mittels einer Deklarierungspflicht). Es gibt aber auch themenspezifische Regulierungsansätze (z.B. zur Bekämpfung von Desinformation – auch – mittels Deepfakes) wie auch Ansätze der Koregulierung oder gänzlicher Selbstregulierung der Plattformen.

## Deepfakes im Journalismus

Als ein wichtiges Element im Umgang mit Deepfakes wird immer wieder das Handeln des Journalismus diskutiert. Die interviewgestützte Analyse von Deepfakes im Journalismus zeigt, dass Redaktionen von Schweizer Medienorganisationen im Herbst/Winter 2022 noch eher selten mit Deepfakes konfrontiert waren. Die Seltenheit von Deepfakes speziell im Schweizer Kontext hatte entsprechend Einfluss auf die Thematisierung in den Ausbildungsorganisationen wie auch auf die Einschätzung durch Vertreterinnen und Vertreter von Medienorganisationen selbst. Mit der zunehmenden Verbreitung und Zugänglichkeit von Technologie zur Erstellung von Deepfakes scheint es aber durchaus realistisch, dass in Zukunft mehr Fälle auftreten werden.

In der journalistischen Praxis werden Deepfakes als technischer Sonderfall von Desinformation betrachtet, die es im Rahmen der Faktenüberprüfung zu erkennen gelte. Aus Sicht der Interviewten hilft beim Umgang mit Deepfakes die Orientierung an grundlegenden journalistischen Normen und Standards, welche in der Ausbildung erlernt und in der Praxis in Erinnerung gerufen und angewendet werden sollten. Dazu gehört auch, dass Informationen aus sozialen Netzwerken immer durch andere Quellen verifiziert werden sollten. In den untersuchten Medienorganisationen gibt es allerdings einen Bedarf an Personen oder spezialisierten Abteilungen, die für komplizierte Fälle der Verifikation von Videos zuständig sind. Einige Redaktionen haben bereits solche Teams oder Personen, um Inhalte aus sozialen Netzwerken oder Videos zu prüfen (Open-Source-Intelligence-Expertinnen und -Experten, Faktenchecker) resp. sind daran, solche Stellen aufzubauen. Aufgrund knapper finanzieller Ressourcen im Journalismus dürfte die Einsetzung oder der Ausbau von internen Verifikationsteams nicht bei

allen Medienorganisationen realistisch sein. Auch ist es fraglich, wie verlässlich technische Tools zur Erkennung von Deepfakes künftig noch sind.

Insgesamt werden Deepfakes von den interviewten Personen überwiegend als Risiko wahrgenommen. Allerdings wird in der journalistischen Praxis kaum ein Unterschied zum umfassenderen Problem der Desinformation gesehen. Die Interviewten befürchten, dass Deepfakes den Vertrauensverlust in Medien und das unbegründete Hinterfragen von Fakten durch einen Teil der Nutzerinnen und Nutzer noch weiter verstärken werden. Neben Verifikationsprozessen und Faktenüberprüfung wird von den interviewten Personen deshalb die Sensibilisierung des Publikums für Deepfakes und die Information über die Verifikationsprozesse als konstruktiver Weg gesehen, um das Vertrauen in die Medien zu fördern. Es scheint also nicht ausreichend, wenn nur aufseiten der Medien Informationen verifiziert werden und ein Bewusstsein für manipulierte Information vorhanden ist, sondern es braucht in der ganzen Gesellschaft ein Bewusstsein für manipulierte Information und für die Notwendigkeit von Quellenkritik, was auf die Relevanz der Medienkompetenzförderung verweist.

Auch in der journalistischen Ausbildung werden Deepfakes als Teil des Umgangs mit Desinformation behandelt. Hauptziele der Ausbildung sind es, angehenden Journalistinnen und Journalisten das Wissen zu vermitteln, wie Informationen gemäss professioneller Normen verifiziert werden können und welche technischen Möglichkeiten zur Manipulation von Quellenmaterial existieren. Dabei werden Deepfakes mehrheitlich als Beispiele in bestehenden Kursen behandelt. Die Orientierung an grundlegenden journalistischen Normen und Standards wird durch die Komplexität der Identifizierung von Deepfakes immer wichtiger. Diese Normen werden in der Ausbildung erlernt und mit Deepfakes in Verbindung gebracht. Aufgrund der Verbesserung und breiteren Zugänglichkeit der Technologie kann davon ausgegangen werden, dass Deepfakes als Thema und Kompetenzen zur Überprüfung komplexer Information in der Ausbildung in Zukunft wichtiger werden. Zudem scheint eine kontinuierliche Weiterbildung von Medienschaffenden in Bezug auf technische Entwicklungen und Verifikationsmöglichkeiten unabdingbar.

## Deepfakes in der Politik

Ein viel diskutiertes Thema ist dasjenige der möglichen politischen Implikationen von Deepfakes. Es bestehen anhaltende Befürchtungen rund um den missbräuchlichen Einsatz von Deepfakes zur Schädigung von Demokratie und Wirt-

schaft. Aus diesem Grund haben wir die mögliche Rolle von Deepfakes in der Schweizer Politik und in der Wirtschaft untersucht.

In unserer Umfrage unter Schweizer Parlamentarierinnen und Parlamentariern und Angehörigen der Bundesverwaltung gab eine Mehrheit der Befragten an, dass Deepfakes entweder bereits im politischen Betrieb diskutiert werden, oder sie äusserten den Wunsch, dass dem Thema mehr Aufmerksamkeit gewidmet werden sollte. In Bezug auf Chancen und Risiken zeigte sich ein eindeutiges Bild: Die Befragten sahen fast ausschliesslich Risiken. Besorgnis besteht insbesondere hinsichtlich der Auswirkungen auf die Schweizer Demokratie und die politischen Institutionen. Bezüglich beider Aspekte sahen die Befragten zudem mehrheitlich eine hohe oder sehr hohe Eintrittswahrscheinlichkeit. Als relevante Risiken wurden ausserdem benannt: das Risiko, dass ein Deepfake über einen selbst kursiert, dass man mittels eines Deepfakes getäuscht wird sowie die aus Deepfakes resultierenden Risiken für internationale Beziehungen. Die Mehrheit der Antworten bringen allerdings zum Ausdruck, dass die Wahrscheinlichkeit des Eintritts derartiger Risiken als eher niedrig bis sehr niedrig einzustufen ist. Die Befragung zeigt schliesslich auch, dass gegenwärtig selten konkrete Schutzmassnahmen gegen Deepfakes ergriffen werden.

Zur Konkretisierung der möglichen Nutzung von Deepfakes in der Schweizer Politik wurden Szenarien erarbeitet. Diese machen systematische Aussagen über mögliche Angreifer bzw. Verursacher, den Angriffstyp sowie den Adressatenkreis. Dabei zeigt sich, dass ein Grossteil der in den Szenarien vorgestellten Angriffsformen bereits mit heute verfügbaren Deepfake-Technologien durchgeführt werden können. Hierzu zählen die Erpressung bzw. Einschüchterung und die Rufschädigung von Politikerinnen und Politikern, Funktionsträgern usw. Auch können Deepfakes dazu beitragen, zum Hass aufzustacheln und zur Gewalt aufzurufen. Zudem eignen sich Deepfakes auch zur Rufschädigung politischer Institutionen und zur Erlangung vertraulicher Informationen, indem Mitarbeitende und ggf. biometrische Authentifizierungssysteme überlistet werden. Die Szenarien befassen sich schliesslich mit den möglichen Effekten auf gesellschaftlicher Ebene. Neben der Beeinflussung von Wahlen und der Beeinflussung von politischen Entscheidungsprozessen werden auch Möglichkeiten zur Verschärfung sozialer Spannungen, Beschädigung der Demokratie, Gefährdung der öffentlichen Sicherheit und Beeinflussung der internationalen Beziehungen erörtert.

## Deepfakes in der Wirtschaft

Sorgen über die zunehmende Nutzung von Deepfakes für wirtschaftskriminelle Zwecke waren Anlass zur Untersuchung solcher Konstellationen in der Schweiz.

Der Medien- und Literaturüberblick zeigt, dass Deepfakes für wirtschaftskriminelle Zwecke bereits heute erfolgreich verwendet werden. Ausgehend von den (technologischen) Einsatzmöglichkeiten von Deepfakes und der Operationsweise krimineller Gruppierungen ist davon auszugehen, dass Deepfakes zunehmend Eingang in das Angriffsrepertoire professioneller krimineller Akteure finden werden. Zur weiteren Konkretisierung wurden auch für den Wirtschaftsbereich Szenarien ausgearbeitet. Diese verdeutlichen Einsatzmöglichkeiten von Deepfakes im Bereich der Wirtschaftskriminalität, insbesondere für Identitätsbetrug und -diebstahl, Rufschädigung von Unternehmen, Onlinewerbebetrug mittels synthetischer Profile bis hin zur Marktmanipulation und Manipulation von demokratischen Entscheidungsprozessen zum Nachteil der Wirtschaft oder von bestimmten Wirtschaftssektoren. Zudem können mit Deepfakes Mitarbeitende getäuscht und biometrische Authentifizierungssysteme überwunden werden, sodass Angreifer Zugang zu gesicherten Systemen erhalten und so z.B. zu Erpressungszwecken Ransomware einspielen oder Wirtschafts- und Industriespionage betreiben können.

Eine Medieninhaltsanalyse zeigte ausserdem, dass Deepfakes jenseits der Risikodebatte in verschiedenen Anwendungsbereichen auch sinnvolle Nutzungsmöglichkeiten bieten. Chancen von Deepfakes ergeben sich demnach vor allem in den Bereichen Unterhaltung, Bildung sowie Werbung. Der Rückgriff auf Deepfake-Technologien in diesen Bereichen kann neue Unterhaltungsangebote schaffen, Fankultur unterstützen und die Kundenbindung stärken, Kosten- und Zeiteinsparungen bewirken, neue satirische Angebote schaffen. Abseits der ökonomischen Verwertungsmöglichkeiten können solche Anwendungen auch den Schutz von Minderheiten und Identitäten ermöglichen, zur Verbesserung von Bildungsangeboten genutzt werden und zur Steigerung der Lernmotivation sowie zur Stärkung der Medienkompetenz beitragen. Mit den unterschiedlichen Chancen können jedoch auch neue Herausforderungen einhergehen, etwa Urheberrechtsfragen.

## Fazit und Empfehlungen

Neue Technologien zur Synthetisierung und Manipulation von Medien werden künftig einen festen Platz in der Alltagskultur einnehmen. Wir haben in dieser Studie diskutiert, wie künftig mit dieser «veränderten medialen Realität» umgegangen werden kann, in der täuschend echte Fälschungen mit originalen Medieninhalten um die Gunst der Rezipienten buhlen. Zunächst verdeutlicht unsere Studie, dass Deepfakes gegenwärtig durch die Mehrheit der Schweizer Bevölkerung, im Bereich des Journalismus und in der Politik als Risiko betrachtet werden. Insbesondere im Journalismus und in der Politik werden Deepfakes auch häufig in den Bereich von Desinformation eingeordnet. Dass Deepfakes mehr sein können, zeigen hingegen unsere Betrachtungen zu Recht und Wirtschaft: Während einige Verwendungen von Deepfake-Technologien tatsächlich in den Bereich des rechtlich Unerlaubten fallen, sind Deepfakes in anderen Fällen von der Meinungs-, Informations- und Kunstfreiheit oder dem Urheberrecht geschützt. Insbesondere beim Blick auf Deepfakes in der Wirtschaft wird deutlich, dass Deepfake-Technologien auch enormes kreatives und ökonomisches Potenzial aufweisen. Der Blick auf die öffentliche Wahrnehmung macht deutlich, wie sehr die gewählte Begrifflichkeit die Richtung des Diskurses bestimmt: Wenn von synthetischen Medien die Rede war, werteten die Befragten den Untersuchungsgegenstand positiver, als wenn von Deepfakes die Rede war.

Die unerwünschten Folgen von Deepfakes lassen sich weder durch einzelne regulatorische noch durch technische Massnahmen verhindern oder eindämmen. Wir schlagen vielmehr vor, dass Schäden durch Deepfakes durch eine Mischung verschiedener Massnahmen auf mehreren Ebenen und durch unterschiedliche Akteure bekämpft und vorgebeugt werden sollten. Gefragt sind Massnahmen durch die Politik, aber auch von Organisationen aus praktisch allen Branchen, weil die Folgen von Deepfakes sehr wahrscheinlich alle Organisationen tangieren werden. Gefragt sind auch Plattformbetreiber, Medienorganisationen und Nachrichtenagenturen sowie Akteure aus der Kommunikationsbranche. Nicht zuletzt sind Bildungseinrichtungen und die breite Gesellschaft gefragt, denn Bürgerinnen und Bürger, die Medieninhalte mit einer gesunden kritischen Distanz bewerten, sind weniger anfällig, in die Irre geführt zu werden.

Wir bewerten es hingegen eher kritisch, wenn zu sehr auf Fortschritte bei der technischen Erkennung von Deepfakes gesetzt wird: Diese Hoffnung besteht seit vielen Jahren und doch ist uns keine Detektionssoftware bekannt, welche die Versprechungen einlösen kann. Dies soll jedoch nicht in Abrede stellen, dass die Weiterentwicklung von Detektions-, aber auch neuer Authentifizierungs- und

Kennzeichnungsverfahren ein Element einer umfassenden Strategie zum Umgang mit Deepfakes sein kann.

In unserer Studie haben wir eine breite Palette an möglichen missbräuchlichen Folgen von Deepfakes betrachtet und Möglichkeiten ihrer Adressierung erörtert. Auf dieser Grundlage formulieren wir folgende *Handlungsempfehlungen*:

**Plattformregulierung:** Eine zentrale Empfehlung der Studie ist das Vorantreiben der staatlichen Bemühungen zur Plattformregulierung. Hier wäre es nicht nur ratsam, eine Verpflichtung der Plattformbetreibenden zur Löschung oder Sperrung von gemeldeten Deepfakes bei begründetem Verdacht auf eine Rechtsverletzung einzuführen. Mit der Verpflichtung der Plattformbetreibenden zur Einrichtung eines Meldesystems rechtswidriger Inhalte inklusive Transparenzvorgaben und Widerspruchsmöglichkeiten würden die Rechte sowohl der Opfer von Deepfakes als auch der Betroffenen unberechtigter Löschungen gestärkt.

**Bildung und Selbstverantwortung der Bürgerinnen und Bürger:** Eine weitere wichtige Empfehlung ist die Stärkung von Bildungsmassnahmen und die Betonung der Selbstverantwortung der Bürgerinnen und Bürger. Denn auch wenn der durch Deepfakes befürchtete flächendeckende Vertrauensverlust in Medieninhalte noch nicht eingetreten ist, wird es angesichts von Deepfakes und generativer KI zunehmend schwieriger, echte Inhalte von Fälschungen zu unterscheiden. Umso wichtiger wird also Medienkompetenz, damit jeder und jede Einzelne diese Differenzierung auch in Zukunft noch vornehmen kann bzw. andere Methoden der Überprüfung der Glaubwürdigkeit eines Inhalts erlernt, wie etwa die Überprüfung der Quelle. Damit Bürgerinnen und Bürger ihre Selbstverantwortung wahrnehmen können, bedarf es allerdings begleitender Massnahmen, insbesondere im Bildungsbereich zur Förderung von Medienkompetenz. Ebenso kann die Stärkung der staatlichen Finanzierung für Opferberatungsstellen, die auf Cyberdelikte spezialisiert sind, einen wichtigen Beitrag leisten, sodass beispielsweise Opfer von Cybermobbing durch Deepfake-Pornografie effektiver Unterstützung erhalten können.

**Vorbereitung auf Deepfakes:** Organisationen in allen gesellschaftlichen Bereichen und in allen Wirtschaftsbranchen sollten sich auf die zunehmende Allgegenwart von Deepfakes vorbereiten. Die Beschäftigung mit den Chancen von KI-basierten Technologien ist bei vielen Akteuren bereits in vollem Gange. Auch der Einsatz von Deepfake-Technologien kann in verschiedenen Bereichen zu Kostensenkungen und anderen ökonomischen Vorteilen beitragen. Andererseits lassen sich Deepfakes gegen beliebige Organisationen auch auf missbräuchliche Weise verwenden. Um sich gegen die cyberkriminelle Nutzung von Deep-

fakes zu wappnen, sollten Organisationen eine interne Risikoabschätzung vornehmen und ggf. zu treffende präventive und reaktive Massnahmen eruieren. Zu diesen können Weiterbildungsmassnahmen, die Schaffung von spezialisierten Krisenreaktionsteams oder das Setzen auf fortschrittliche Authentifizierungsmassnahmen zählen.

**Hochhalten journalistischer Standards:** Schliesslich kann auch der Journalismus einen wichtigen Beitrag für eine Zukunft der verantwortungsvollen Nutzung von Deepfake-Technologien leisten: Das Hochhalten journalistischer Standards kann der Erkennung von Deepfakes und der unmittelbaren Aufklärung der Bevölkerung über gefälschte Inhalte dienen. Dadurch kann der Journalismus seine gesellschaftliche Funktion in Form der Bereitstellung nach professionellen Standards produzierter Inhalte auf Basis geprüfter Informationen besser erfüllen.

Angesichts der enormen Geschwindigkeit der technologischen Veränderung ist aber damit zu rechnen, dass die hier gemachten Vorschläge schon bald von der Realität überholt werden. Die von uns adressierten Akteure aus Politik und anderen Bereichen sollten sich dieser Dynamik bewusst und darauf eingestellt sein, neue Entwicklungen zu erkennen und mit adäquaten weiteren Schritten und Massnahmen zu reagieren.

# Summary

A deepfake is an audio or (moving) image content synthesised or manipulated with the aid of artificial intelligence technologies, which appears to be authentic and in which a person says or does something that he or she has never said or done. Since 2017, when synthetic and manipulated media content was first referred to as “deepfake”, the term has become firmly entrenched in political and media debate.

Several years after its introduction, we can now draw some mixed conclusions: some political deepfakes and deepfake-based fraud cases, for example nuisance or scam calls, appear to confirm certain concerns. On the other hand, a variety of useful applications have been created that are based on synthetic and manipulated media content. Neither the widely feared large-scale use of deepfakes in disinformation campaigns nor a major information apocalypse have materialised to date.

The purpose of this study is to assess the opportunities and risks associated with deepfakes for Switzerland. In view of the progress that has been achieved in the field of deepfake technologies, the first question to be examined concerns the *current status of these technologies*. The question of *how much people know about deepfakes* and how they assess the associated risks and opportunities, has been barely researched to date. The study also looks into the question of how the use of deepfakes has to be assessed *at the legal level*: what is permitted, and where do the boundaries have to be drawn? And is the existing legislation sufficient, for example in order to protect the population against the unlawful use of deepfakes? One of the key factors that is frequently discussed in the context of dealing with deepfakes concerns the responsible-minded approach on the part of *journalists*. How do Swiss journalists perceive deepfakes and how do they deal with them? Another issue that is frequently discussed concerns the potential *political implications* of deepfakes. Here we ask ourselves which role deepfakes could play in politics and which options for action exist to enable policy-makers and other players to avoid undesirable impacts. Recently, a great deal of potential has been attributed to the use of deepfakes for *financial purposes*. In this context, our study examines the potential economic opportunities that could arise through the use of deepfake technologies, as well the improper economic purposes for which these could be used. The study also investigates the possible range of preventive and reactive protection measures that could be

taken by (potentially) affected parties. Finally, our study pursues the following two objectives: to provide information about the opportunities and risks associated with deepfake technologies, and to draw attention to available *courses for action* that facilitate a possible responsible approach to dealing with deepfakes so that the perception of the opportunities can remain intact, while the risks can be at least mitigated, if they cannot be fully avoided. The study examines in detail the issues referred to above in separate chapters. The key findings are summarised below under the following headings:

- Deepfake technologies – analysis of current status and trends
- Awareness of deepfakes among the population
- Deepfakes and the law
- Deepfakes in journalism
- Deepfakes in politics
- Deepfakes in the economy
- Conclusions and recommendations

## Deepfake technologies – analysis of current status and trends

The examination of existing and foreseeable technologies for the creation of deepfakes shows that significant progress has been made in the past few years with regard to the AI-based synthesisation and manipulation of image, audio and text content.

There are currently five methods for creating *image-based* deepfakes:

- *Facial re-enactment:* manipulation of facial expressions
- *Face morphing:* blending of two or more faces
- *Face swapping:* exchanging or substituting one face by another
- *Face generation:* creation of faces that do not exist
- *Full body puppetry:* alteration of the pose or movements of an entire body or part thereof.

In the field of *audio* deepfakes, in the past few years the trend has moved in the direction of increasingly credible results, which can now also be attained with a reduced need for training material.

Thus it is already possible to create a broad range of deepfake content today. With modern deepfake technologies it is possible to emulate a person's manner of speaking, facial expressions and even body movements, always assuming that the creators possess sufficient input data (for example, speech samples or face photos). With modern text generators it is also possible to imitate a person's writing style and language, which can be used as input for deepfakes. In addition, backgrounds can be fully synthesised and modified at ever higher quality. This means that, in addition to deepfakes depicting people, it is also possible to create realistic scenarios showing fake plane crashes or natural disasters. Today, it is possible to create high-quality audio deepfakes with comparatively little outlay. Similarly, AI image generators can be used to produce deceptively realistic deepfake images of people with whom the generator is familiar, i.e. public figures in particular. By training generators with supplementary material or by adding own images to a text command, it is also possible to generate images of any other persons as desired.

Trials we carried out ourselves aimed at generating deepfake videos demonstrate that, despite the enormous technological progress that has been made, a great deal of technical expertise and vast computing capacities, as well as time and financial resources, are still required in order to create credible video deepfakes. We therefore agree with the assessment which assumes that, there are currently no indications that the general population will be able to generate deceptively realistic deepfake videos and thus flood the media landscape with them. This situation could change if existing deepfake generators were to be made simpler to use, or if text-to-video processes were to become more advanced and readily available to the general public. With text-to-video technologies it would be possible to generate deepfake videos by entering text commands, in the same way as with existing AI-based text or image generators such as ChatGPT or Midjourney. It may be assumed that the best known programs of this type will implement safeguards in order to prevent harmful or unethical usage. On the other hand, such safeguards can of course be circumvented. And obviously, other generators could be developed that waive ethical guidelines. So even if there is no immediate danger that credible deepfake videos will flood the (social) media landscape, their generation will become simpler in the future. Based on these findings, we anticipate that, in the near future, deepfake videos will primarily be produced by individuals or groups of players who possess the necessary expertise and computing capacities.

But as we shall see below, it is not only large quantities of deepfakes that can have harmful effects. In fact, just one single deepfake can give rise to major harm at the political, legal, societal or economic level. And as the reactions to a manipulated photo of the Pope in a down-filled jacket have shown (in 2023), there does not necessarily have to be an intent to deceive or cause harm. People can also be deceived by media content that was created for entertainment purposes.

Discussions of technical remedial measures include methods for authenticating original content, designating deepfake content and identifying deepfake videos. But all these measures have certain weaknesses. The fact that they can be readily circumvented indicates that the authentication of credible content and the designation of deepfake content cannot be trusted. In addition to causing major harm *per se*, deepfakes falsely signed as trustworthy could also fundamentally damage trust in signed documents. It is presumably well-resourced (state) entities that are able to circumvent technical safeguards. But this could at the same time mean that the majority of circulating deepfakes would be designated as such and the majority of trustworthy content would be signed and identifiable as such. In combination with other measures, signing could – despite its weaknesses – thus become an element of an overall strategy for suppressing undesirable impacts of deepfakes.

Our evaluation of the status of the existing literature and the results of our own tests have shown that readily available deepfake detectors are currently unreliable. And estimates according to which the generators are constantly a step ahead of the detectors indicate that improvements are not to be anticipated. However, in view of the investments in the development of detectors and the technical challenges facing *GANs* (generative adversarial networks) and other deepfake technologies, the possibility nonetheless exists that the accuracy of future processes could increase.

## Awareness of deepfakes among the Swiss population

The attitude of the general public towards deepfakes has barely been researched to date. With the aid of a representative survey we wanted to find out how deepfakes are perceived by the Swiss population. We asked the participants to tell us about their experiences with deepfakes; how they assess the risks and op-

portunities of deepfake technologies; and whether they are able to distinguish between deepfakes and genuine videos.

As of September 2023, many people were still unfamiliar with the term “deepfakes”. Only slightly more than half the respondents had already heard of them, while almost fifty percent stated they had already seen some. Only a tiny majority had already had experience with the generation of deepfakes (two percent) and distribution of deepfakes (three percent). On the one hand, this means that people’s awareness of this issue is primarily based on reports in the media, not on their own experiences. But this also opens up the opportunity to provide education and information regarding the meaningful application of deepfake technologies.

Among the Swiss population, deepfakes are strongly associated with risks, and this applies to almost all socio-demographic groups. However, the assumed effect of deepfakes plays an influencing role on the opinion of third parties. The greater the assumption that other people are influenced by deepfakes, the higher the assessment of the risk posed by deepfakes for society as a whole. Thus there appears to be a somewhat vague notion regarding the impacts of deepfakes on society, which increases the level of risk perception. By contrast, using the term “synthetic media” instead of “deepfake” has a positive effect on the perception of the opportunities associated with deepfake technologies. On balance, the findings indicate that there remains a great deal of uncertainty among the population regarding deepfakes. One relevant finding arising from the study concerns the fact that women assess the risk to themselves – i.e. the threat to their private sphere or the possibility of becoming a victim of a deepfake – more highly than men do.

Our experiments show that people in Switzerland are barely able to distinguish between deepfakes and genuine videos, especially when the former are of good to very good quality. This means that a high degree of media literacy is of particular importance. Our study shows that *social media literacy*, i.e. skill in the use of the new media on the Internet, has a positive effect on the ability to identify deepfakes. By contrast, general media literacy and Internet skills do not have this effect. Our brief literacy intervention, i.e. the provision of technical assistance in identifying deepfakes immediately prior to showing the videos, did not have any effect in terms of identification capability. There was also no *backfire effect*. So our literacy intervention did not cause the participants to become overcritical, and consequently they even considered real videos to be deepfakes.

## Deepfakes and the law

The law has a dual role to play with respect to deepfakes: it not only has to protect them as a form of artistic creation or expression of opinion, but also has the important duty of providing protection against harmful and/or undesirable impacts of deepfakes, in particular through the implementation of criminal law, copyright protection and protection of privacy and data, as well as through public law (media) regulation. This means that deepfakes and/or their initiators are entitled to protection of freedom of opinion, information and artistic expression, because even untrue, fake or deceptive information is to a certain extent protected under fundamental rights. In addition, in certain circumstances deepfakes can also be protected under copyright law. This protection is not absolute, however: fundamental rights can be restricted under certain circumstances, specifically to protect the rights of third parties. For this reason, the main focus of the legal analysis in this study is on protection against harmful usage and impacts of deepfakes.

In Switzerland there are no specific regulations governing deepfakes. Instead, different general legal provisions apply depending on the intended use. “Deepfakes” and other designations used within this context are not official legal terms, but rather are to some extent covered by existing concepts. A distinction has to be made between civil law and criminal law provisions and their procedural enforcement, and public law requirements.

Many “undesirable” uses of deepfakes are covered by existing legislation. Violations of personal rights can be dealt with under civil law. Furthermore, the production of deepfakes normally constitutes an infringement of copyright and data protection entitlements. Some deepfake use cases also fall under criminal law, including for example defamation, extortion or fraud, pornography, sexual harassment and child pornography. It is currently unclear how the recently adopted crime of identity theft will be dealt with in the context of deepfakes. There are still grey areas, for example regarding the classification under criminal law of “embarrassing” or “explicit” deepfakes as pornography or harassment. It is also apparent that in the case of use of deepfakes as an image-based form of sexualised violence, the unlawfulness based on defamation in combination with the right to sexual self-determination is not covered by the existing provisions of the penal code.

The biggest challenges arise in the context of enforcement of entitlements under civil and criminal law. In the Internet, perpetrators are frequently anonymous

or based abroad. Furthermore, initiating proceedings against platforms is associated with procedural uncertainties. In view of the rapid and potentially widespread distribution of content on the Internet, it is almost impossible to halt the spread of a deepfake with the aid of legal measures once it has been put into circulation. Also, the procedural costs associated with civil proceedings can deter people affected by deepfakes from taking legal action.

Deepfakes can also play a role in the context of legal proceedings, for example as evidence in criminal proceedings. However, the presentation of forged documents in court is by no means a recent development. For many years, software such as Photoshop has been facilitating the manipulation of texts and images. Deepfake technology is a new variant that simplifies the production of fake audio and videos, but does not give rise to fundamentally new issues. The existing legal provisions for dealing with forged evidence are applicable.

It is also conceivable that law enforcement authorities could use deepfake technologies in order to solve crimes. Here, for example, it is possible to create a virtual crime scene or reconstruct the sequence of events using videos from mobile phones and surveillance cameras, or body scans, with the aid of deepfake technology. Another hotly debated issue is whether undercover agents may use computer-generated child pornography in order to gain access to online forums that require users to upload their own images or videos. The latter is above all problematic because, under existing legislation, the prosecuting authorities are not permitted to commit one crime in order to solve another (and in our view, child pornography produced with the aid of deepfakes indeed falls under the definition of child pornography). On the other hand, the former is associated with various procedural uncertainties, including in particular the question of how the guarantee of participation rights can be secured during all stages of the proceedings, and hence during the “virtual crime scene investigation” itself, how the resulting evidence can be substantiated and how – and in what form – the digitally collected evidence can be subsequently archived.

Finally, there is the question concerning the regulation of deepfakes: some countries such as the US pursue a technology-based approach that specifically regulates deepfakes, for example in the form of a declaration requirement. There are also other thematic regulatory approaches in some fields related to deepfakes, for example regulation aimed at combating deepfake-based disinformation, but also approaches such as joint regulation or full self-regulation of platforms.

## Deepfakes in journalism

One of the key elements that is frequently discussed in the context of dealing with deepfakes concerns the approach on the part of *journalists*. The interview-based analysis of deepfakes in journalism shows that the editorial offices of Swiss media organisations were rarely confronted with deepfakes in autumn/winter 2022. The rarity of deepfakes in the Swiss context in particular had a corresponding influence on the inclusion of the topic in educational institutions, as well as on the assessment of deepfakes by representatives of media organisations. But with the increasing distribution and availability of technologies for producing deepfakes, it appears highly likely that more cases will arise in the future.

In journalistic practice, deepfakes are regarded as a special case of technical disinformation that has to be checked in terms of factual correctness. In the view of the interviewees, dealing with deepfakes can be supported by observing basic journalistic rules and standards that are learnt during training and have to be recalled and applied in practice. This includes the need to always use refer to other sources in order to verify information that is obtained via social networks. In the participating media organisations, however, there is a need for personnel or specialised sections that are responsible for complicated cases of verification of videos. Some editorial offices already have such teams or personnel whose task is to verify content from social networks or videos (open source intelligence specialists, fact checkers), or are preparing to introduce them. But due to scarce financial resources in the journalism sector, the introduction or expansion of internal verification teams is unlikely to be feasible in all media organisations. There are also uncertainties regarding the long-term reliability of technical tools for identifying deepfakes.

The majority of the respondents perceive deepfakes as a risk. However, in journalistic practice little distinction is made between deepfakes and the more comprehensive problem of disinformation. The respondents fear that deepfakes will intensify the loss of confidence in the media and increase the unfounded scrutinisation of facts by some users. Alongside verification processes and fact checking, the interviewees regard the sensitisation of the public to deepfakes and providing information about verification processes as a constructive way to foster confidence in the media. It thus does not appear sufficient for information to only be verified by the media and that the latter alone should have an awareness for manipulated information. Instead, what is required is an awareness for manipulated information throughout the whole of society and for the necessity for verification of sources, which points to the relevance of the promotion of media skills.

In journalistic training, too, deepfakes are regarded as part of the handling of disinformation. The main objective of the training is to teach prospective journalists how information can be verified in accordance with professional standards, and which technical options exist for manipulating source material. Deepfakes are mostly used as examples in existing courses. Due to the complexity associated with the identification of deepfakes it is becoming increasingly important to focus on basic journalistic rules and standards, which are taught in training courses and brought into association with deepfakes. In view of the improvement of – and more widespread access to – the various technologies, it may be assumed that deepfakes will become more important in journalistic training in the future as a specific topic and in order to improve expertise in the verification of complex information. In addition, the further education of media representatives in the area of technological development and verification methods appears to be essential.

## Deepfakes in politics

The potential political implications of deepfakes is a frequently discussed issue. Fears persist regarding the improper use of deepfakes to harm democracy and the economy. In view of this we examined the potential role of deepfakes in Swiss politics, and subsequently in the economy.

In our survey of Swiss members of parliament and players within the federal administration, the majority of respondents either stated that deepfakes are already being discussed at the political level, or that greater attention needs to be paid to them. The response to the question regarding risks and opportunities was unequivocal: the respondents overwhelmingly view deepfakes as a risk. They express concerns with respect to the impacts on Swiss democracy as well as on the country's political institutions. In both cases the majority of the respondents envisage a high to very high probability of occurrence. The cited relevant risks include: the possibility that they themselves could be affected by a deepfake, that they could personally be deceived by a deepfake and that deepfakes could harm international relations. However, most of the respondents stated that the likelihood of the occurrence of such risks can be regarded as low to very low. The survey also shows that very few specific measures have been taken to date to protect against deepfakes.

Scenarios were drawn up to illustrate the potential use of deepfakes in Swiss politics. These scenarios systematically describe potential attackers or producers of deepfakes, plus the type of attack and the target group. It is apparent

that most of the described types of attack can already be carried out with the aid of currently available deepfake technologies. These include extortion or intimidation and harm to the reputation of politicians, officials, etc. In addition, deepfakes can also be used to incite hate and violence. They can also be used for harming the reputation of political institutions and in order to gain access to confidential information by deceiving personnel and circumventing biometric authentication systems. The various scenarios focus solely on potential impacts at the societal level. In addition to influencing elections and political decision-making processes, discussions include the possibilities of intensifying social tensions, harming democracy, threatening public security and influencing international relations.

## Deepfakes in the economy

Deepfakes are being used to an increasing extent for the commission of financial crime, for example identity fraud and theft. An increase in cases of this nature prompted the implementation of a study of such constellations in Switzerland.

As the overview of media and literature shows, deepfakes are already being successfully used for the commission of financial crime. Based on the potential (technological) uses of deepfakes and the modes of operation of perpetrators it is to be assumed that use of deepfakes will become more widespread among professional criminals. For further clarification, scenarios were drawn up in this area too. They illustrate the potential uses of deepfakes in the area of financial crime, in particular for identity fraud and theft, damage to companies' reputation, online advertising fraud with the aid of synthetic profiles, through to market manipulation and interference with democratic decision-making processes to the detriment of the economy as a whole or certain sectors. In addition, deepfakes can be used to deceive personnel and overcome biometric authentication systems, so that attackers can gain access to secured systems and thus install ransomware for extortion purposes, or carry out business and industrial espionage.

An analysis of media content also showed that, beyond the risk debate, deepfakes can offer meaningful usage potentials in various areas of application. Opportunities arise in particular in the areas of entertainment, education and advertising. In these areas, the use of deepfake technologies can result in the creation of entertainment opportunities, support fan culture, strengthen customer loyalty, save time and costs, and create new satirical products. Apart from opening up economic opportunities, such applications can also facilitate the protection of minorities and identities, improve educational programmes and increase motiva-

tion for learning, as well as enhance media skills. But the various opportunities can also go hand in hand with new challenges, such as copyright issues.

## Conclusions and recommendations

New technologies for synthesising and manipulating media will gain a firm footing in our everyday culture. In this study we discuss how this “changed media reality” can be approached, in which convincing falsifications use original media content to deceive recipients. Initially, our study illustrates that deepfakes are regarded as a risk by the majority of the Swiss population, and by journalists and politicians. In the areas of journalism and politics in particular, deepfakes are commonly classified as disinformation. But the fact that they can be more than this is underscored by our findings in the fields of legislation and finance. While some applications of deepfake technology are in fact unlawful, in other cases deepfakes are protected in accordance with freedoms of expression and information, or artistic freedom. With regard to deepfakes in the economy, it is clear that these technologies also possess enormous creative and financial potential. And with respect to public perception it is apparent that the selected terminology greatly determines the direction of discourse: when the term “synthetic media” is used, the respondents assess the research topic in a much more positive light than when reference is made to “deepfakes”.

The undesirable consequences of deepfakes cannot be prevented or lessened either through specific regulatory or technical measures. Instead, we propose that harm caused by deepfakes should be combated and prevented through the implementation of a combination of various measures at several levels and by a variety of players. Measures are required at the political level, but also on the part of organisations from practically every sector, because the consequences of deepfakes are very likely to affect all organisations. The involvement of platform operators, media organisations and press agencies, as well as players from the communications sector, is also necessary. The same applies with respect to educational institutions and society as a whole, because people who assess media content calmly and critically are less likely to be deceived.

By contrast, in our view it is perhaps unwise to count too strongly on the development of tools for identifying deepfakes: this hope has existed for many years, but we are not aware of any software that is able to fulfil it. This does not mean that the further development of detection, authentication and identification pro-

cesses cannot become a component of a comprehensive strategy for dealing with deepfakes.

In our study we observed a broad range of potential consequences of the improper use of deepfakes, and on this basis, compiled *a variety of possible actions* as follows:

**Regulation of platforms:** One of the study's main recommendations concerns continued efforts on the part of the authorities to regulate the various platforms. Here it would be strongly advisable to introduce an obligation on the part of platform operators to delete or block reported deepfakes when there is a well-founded suspicion of a violation of law. By obliging platform operators to install a system for reporting unlawful content, including transparency requirements and appeal options, it would be possible to strengthen the rights of victims of deepfakes as well as of persons whose posts have been deleted without justification.

**Education and personal responsibility:** Another strong recommendation concerns the fostering of educational measures and emphasising the need for citizens to take personal responsibility. Even if the feared comprehensive loss of confidence in media content due to the use of deepfakes has not yet occurred, it will become increasingly difficult to distinguish between genuine and fake content with the widespread use of deepfakes and generative AI. Media skills will become increasingly important, so that everyone will be able to distinguish between genuine and fake content in the future, and be able to apply other methods for determining the credibility of a given content, for example by verifying the source. In order for people to take personal responsibility, however, various support measures will be required, in particular in the field of education aimed at improving media skills. Similarly, increased financing for victim advisory centres specialising in cyber crime can make a significant contribution, so that victims of cyber mobbing in the form of deepfake pornography can be offered effective support.

**Preparation for deepfakes:** Organisations in society as a whole and in all sectors of the economy should prepare themselves for the increasing prevalence of deepfakes. Numerous players are already investigating the opportunities associated with AI-based technologies. And the use of deepfake technologies has the potential to cut costs and yield financial benefits in a broad variety of sectors. On the other hand, deepfakes can also be used improperly against any desired organisation. In order to defend themselves against the unlawful use of deepfakes, organisations should carry out an internal risk assessment and where necessary identify appropriate preventive and reactive measures. This can include further

education measures, the creation of specialised crisis response teams or the implementation of sophisticated authentication measures.

**Upholding journalistic standards:** Journalism can also make a valuable contribution towards the future responsible use of deepfake technologies: upholding journalistic standards can support the identification of deepfakes as well as facilitate the direct provision to the public of information about fake content. In addition, journalism can also better perform its societal function by providing content produced in accordance with professional standards on the basis of verified information.

In view of the extremely rapid pace of technological development, it has to be assumed that the proposals described here will soon be overtaken by reality. The various players approached by us in the framework of this study need to be aware of this dynamic, be prepared to identify new developments and respond to them with adequate further steps and measures.

# Résumé

Un deepfake, ou hypertrucage, est un contenu audio ou visuel (animé) synthétique ou manipulé à l'aide de techniques d'IA, qui semble authentique et qui, la plupart du temps, fait faire ou dire à un individu quelque chose qu'il n'a jamais fait ou dit. Depuis que ces contenus médiatiques synthétiques et manipulés ont été qualifiés pour la première fois de « deepfake » en 2017, ce terme a pris une place de choix dans le débat politique et médiatique.

Plusieurs années après leur apparition, le bilan est mitigé. D'un côté, de multiples deepfakes politiques et arnaques basées sur des hypertrucages, comme récemment les faux appels choc, semblent confirmer certaines craintes. D'un autre côté, on a vu apparaître un certain nombre d'applications utiles qui s'appuient sur ces contenus médiatiques synthétiques et manipulés. Néanmoins, ni l'utilisation massive de deepfakes dans le cadre de campagnes de désinformation, ni l'« apocalypse de l'information », tant redoutées, ne se sont encore produites.

Notre étude a pour objectif d'évaluer les opportunités et les risques que présentent les deepfakes pour la Suisse. Compte tenu des progrès réalisés dans ce domaine, nous avons commencé par dresser un *état des lieux technologique*. Nous avons aussi complété les recherches, encore rares à ce jour, sur *ce que le grand public sait des deepfakes* et sur sa perception de leurs opportunités et de leurs risques. L'étude examine aussi l'utilisation des deepfakes sous l'angle *juridique* : Qu'est-ce qui est permis et où sont les limites ? La législation actuelle suffit-elle notamment à protéger la population contre leur utilisation abusive ? En matière d'hypertrucages, la responsabilité des *journalistes* est un des éléments importants qui revient régulièrement dans le débat. Mais comment les deepfakes sont-ils perçus dans la profession en Suisse, et que font les journalistes face à eux ? La question des éventuelles *implications politiques* des deepfakes fait également l'objet de vives discussions. Nous nous interrogeons ici concrètement sur le rôle que les hypertrucages pourraient jouer dans la politique et sur les options dont disposent les responsables politiques et autres acteurs pour en éviter des effets indésirables. Depuis peu, le potentiel des deepfakes dans le domaine *économique* se révèle lui aussi très élevé. Nous examinons donc à la fois les opportunités économiques que présentent les technologies deepfake, les buts économiques abusifs qu'elles peuvent viser et les mesures de protection préventives et réactives à disposition des personnes (potentiellement) concernées.

Enfin, notre étude n'a pas pour seul but de fournir une base de connaissances sur les opportunités et les risques des technologies deepfake, mais aussi de *formuler des recommandations* sur la gestion responsable des deepfakes, de manière à continuer d'en percevoir les opportunités tout en évitant les risques, ou du moins en les limitant. Cette étude approfondit ces questions et présente en résumé ses principales conclusions dans différents chapitres :

- Analyse de la situation actuelle et des tendances des technologies deepfake
- Perception des deepfakes par la population
- Les deepfakes et le droit
- Les deepfakes et le journalisme
- Les deepfakes et la politique
- Les deepfakes et l'économie
- Recommandations

## Analyse de la situation actuelle et des tendances des technologies deepfake

La présente étude dresse un état des lieux de la recherche sur les technologies deepfake à l'appui d'une analyse de la situation et des tendances. Nous avons choisi de nous concentrer sur les hypertrucages basés sur l'image et le son. En premier lieu, nous nous sommes penchés sur les moyens techniques de création de deepfakes qui existent et sur ceux qui se profilent déjà. Dans un deuxième temps, nous avons recensé les procédés permettant de distinguer les contenus authentiques des contenus synthétiques ou manipulés.

Comme le montre l'analyse des techniques de création de deepfakes existantes et prédictibles, des progrès considérables ont été réalisés ces dernières années en matière de synthèse et de manipulation de contenus visuels, audio et textuels basées sur l'IA.

Dans le domaine des deepfakes *basés sur l'image*, il existe actuellement cinq procédés :

- *facial reenactment* : manipulation des expressions faciales,
- *face morphing* : fusion de plusieurs visages,

- *face swapping* : échange ou remplacement d'un visage par un autre,
- *face generation* : création de visages qui n'existent pas dans la réalité, et
- *full body puppetry* : modification de la pose ou des mouvements d'une partie ou de l'ensemble du corps (comme une marionnette).

Dans le domaine des deepfakes *audio*, l'évolution technologique de ces dernières années a permis d'obtenir des résultats toujours plus crédibles et de le faire avec toujours moins de matériel d'entraînement.

Il est donc déjà possible aujourd'hui de créer une large palette de contenus synthétiques et manipulés. Les technologies deepfake modernes permettent de copier la façon de parler, les expressions du visage et même les mouvements corporels d'une personne – à condition de disposer en permanence d'une puissance de calcul et de données d'entrée suffisantes. S'y ajoute le fait que les générateurs de texte modernes imitant le style et le langage d'un individu sont également capables de produire de telles entrées. Quant aux arrière-plans, ils peuvent être entièrement synthétisés et modifiés avec une qualité toujours meilleure. Cela permet non seulement de créer des deepfakes où figurent des personnes, mais aussi des scénarios réalistes comme un faux crash d'avion ou une catastrophe naturelle. Fabriquer des hypertrucages audio de haute qualité à un coût relativement faible est déjà une réalité. De même, un générateur d'images d'IA qui dispose déjà de données sur une personne peut produire des visuels deepfake trompeusement réalistes de celle-ci. En entraînant les générateurs avec d'autres matériaux ou en ajoutant ses propres illustrations à une commande textuelle, il devient imaginable de produire des images de n'importe qui.

Toutefois, les expériences que nous avons nous-mêmes menées pour créer des vidéos deepfake confirment que, malgré les énormes progrès de la technologie, un grand savoir-faire technique et de gigantesques capacités de calcul, c'est-à-dire beaucoup de temps et d'argent, sont encore nécessaires pour obtenir une qualité crédible. Nous souscrivons donc aux évaluations suggérant qu'il n'existe actuellement aucun signe indiquant que les gens soient capables de créer des vidéos deepfake trompeuses en masse et d'en « inonder » le paysage médiatique. Cette situation pourrait changer s'il devient plus facile d'utiliser des générateurs de deepfake vidéo ou si les procédés de *text to video* gagnent en maturité et deviennent accessibles à de larges groupes de population. De tels outils, comme les générateurs de texte ou d'images actuels basés sur l'IA tels que ChatGPT ou Midjourney, permettraient de créer à l'envi des vidéos deepfake par le biais de commandes textuelles. Il est raisonnable de supposer que

ces IA les plus connues mettront en place des protections pour empêcher les utilisations dommageables et contraires à l'éthique, même si toute mesure de sécurité peut être contournée – et même s'il est imaginable que d'autres générateurs se passent de directives éthiques à l'avenir. C'est pourquoi, si le risque que les médias (sociaux) soient « inondés » de vidéos deepfake crédibles n'est pas imminent, il sera toujours plus aisé d'en créer au fil du temps. À la lumière de ces résultats, nous estimons que, dans un avenir proche, seuls les acteurs individuels ou organisés disposant du savoir-faire et des capacités de calcul techniques nécessaires seront en mesure de produire des vidéos synthétiques ou manipulées.

Dans les chapitres suivants, nous verrons cependant que les hypertrucages peuvent avoir des conséquences néfastes même s'ils ne sont pas nombreux : un seul deepfake est déjà capable de causer des dommages importants de nature politique, juridique, sociale ou économique. Et comme l'ont démontré les réactions à une photo manipulée du pape en doudoune en 2023, il n'est même pas nécessaire d'avoir l'intention de tromper ou de nuire. Les gens peuvent aussi être induits en erreur par des contenus médiatiques créés à des fins purement divertissantes.

Parmi les mesures techniques envisagées pour y remédier, on trouve notamment les procédés d'authentification des contenus originaux, de marquage des contenus deepfake et de détection des vidéos deepfake. Toutes ces mesures présentent des faiblesses et leur surmontabilité technique fondamentale va à l'encontre de l'authentification des deepfakes dignes de confiance et du marquage des contenus deepfake. Quant aux hypertrucages munis d'une fausse signature considérée comme fiable, ils sont non seulement susceptibles de causer eux-mêmes de gros dégâts, mais aussi de porter un préjudice fatal à la confiance que l'on accorde aux documents signés. Les acteurs (étatiques) les plus dotés en ressources sont vraisemblablement les mieux armés pour contourner les mesures techniques de protection. Toutefois, cela pourrait aussi signifier que la masse des deepfakes en circulation serait identifiée, et que la masse des contenus dignes de confiance serait signée comme telle et identifiable. Malgré ses faiblesses, la signature pourrait donc constituer, en combinaison avec d'autres mesures, un élément d'une stratégie globale visant à endiguer les effets indésirables des deepfakes.

Notre analyse de la littérature existante et nos propres observations montrent que, à l'heure actuelle, les détecteurs d'hypertrucages disponibles en libre accès ne sont pas fiables. Selon toute vraisemblance, les générateurs auront toujours une longueur d'avance sur les détecteurs. Toutefois, au vu des investissements

réalisés pour développer des détecteurs et compte tenu des défis techniques auxquels sont confrontés les réseaux adverses génératifs (Generative Adversarial Networks, *GAN*) et d'autres technologies deepfake, la précision des futurs procédés de détection pourrait bien augmenter malgré tout.

## Les deepfakes et la population suisse

L'attitude de la population à l'égard des deepfakes n'a guère fait l'objet de recherches à ce jour. Un sondage représentatif nous a permis d'étudier la manière dont ils sont perçus par les citoyennes et citoyens suisses. Nous avons recueilli leurs expériences en la matière, leur perception des opportunités et des risques de cette technologie, et leur capacité à distinguer les deepfakes des vidéos authentiques.

Les deepfakes sont un concept encore méconnu. À peine plus de la moitié des personnes interrogées en ont déjà entendu parler, tandis que près de la moitié ont déclaré en avoir déjà vu. Seule une très petite minorité a déjà fait l'expérience de la création (2 %) ou de la diffusion de deepfakes (3 %). Cela signifie notamment que le sujet est avant tout perçu en fonction de sa couverture médiatique et non d'une expérience personnelle. Mais cela ouvre également la possibilité de proposer des formations et de communiquer sur le sujet afin de promouvoir une utilisation judicieuse de la technologie deepfake.

Pour la population suisse, les hypertrucages sont fortement associés aux risques, avec peu de différences entre les groupes sociodémographiques. Toutefois, l'effet supposé des deepfakes sur l'opinion de tiers joue un rôle : plus les gens pensent que d'autres personnes sont influencées par des deepfakes, plus ils estiment que leurs risques pour la société sont élevés. Il semble donc que lorsqu'il existe une idée plutôt floue des effets des deepfakes sur la société, cela renforce la perception de leurs risques. En revanche, l'emploi du terme « média synthétique » plutôt que celui de « deepfake » a un impact positif sur la manière dont les opportunités offertes par cette technologie sont perçues. Dans l'ensemble, les résultats indiquent qu'il existe encore beaucoup d'incertitudes au sein de la population à l'égard des deepfakes. Une autre constatation intéressante est que la perception des risques individuels – risque d'atteinte à la vie privée ou d'être soi-même victime d'un deepfake – est plus élevée chez les femmes que chez les hommes.

Selon nos observations, le public suisse a du mal à distinguer les deepfakes des vidéos authentiques, surtout s'ils sont de bonne, voire de très bonne qualité. L'éducation aux médias, ou littératie médiatique, revêt d'autant plus d'importance. Notre étude montre que la *social media literacy*, c'est-à-dire l'aptitude à utiliser les nouveaux médias sociaux sur Internet, a un effet positif sur la capacité à détecter les deepfakes. En revanche, les compétences médiatiques en général et le savoir-faire sur Internet n'ont aucune incidence. Nous avons tenté une petite expérience de littératie en diffusant, juste avant le visionnage de vidéos authentiques ou manipulées, une aide à la reconnaissance des deepfakes. Si cela n'a eu aucun impact sur l'aptitude des personnes interrogées à reconnaitre le vrai du faux, il n'y a pas non plus d'effet de « retour de flamme » (*backfire effect)*. Notre expérience de littératie n'a donc pas non plus induit d'excès de scepticisme à l'égard des vidéos, ni une tendance à considérer toutes les vidéos comme des deepfakes.

## Les deepfakes et le droit

En matière d'hypertrucages, le droit a un double rôle à jouer : d'une part, il protège les deepfakes en tant qu'expression d'un art ou d'une opinion, et d'autre part, il offre une protection contre leurs effets préjudiciables et/ou indésirables, notamment par le biais du droit pénal, de la protection des droits d'auteur et de la protection de la personnalité et des données, ainsi que par le biais de la réglementation (des médias) de droit public. Quiconque crée un deepfake peut donc faire valoir un droit à la liberté d'expression, à la liberté d'information et à la liberté artistique puisque même les déclarations fausses, mensongères ou trompeuses bénéficient, au moins partiellement, de la protection des droits fondamentaux. Cette protection n'est quand même pas absolue : selon les conditions, des restrictions aux droits fondamentaux sont possibles, notamment pour préserver ceux des tiers. C'est pourquoi l'analyse juridique de la présente étude met l'accent sur les dispositions légales de protection contre les applications préjudiciables des deepfakes et leurs effets néfastes.

En Suisse, aucune réglementation spécifique sur les deepfakes n'existe encore. Cependant, en fonction de l'application envisagée, différentes dispositions légales d'ordre général doivent être respectées. Le terme « deepfake », tout comme les autres expressions employées dans ce contexte, n'est pas un concept juridique, mais se retrouve partiellement dans des concepts existants. Il

faut veiller à la distinction entre les dispositions de droit civil et pénal, y compris le droit de procédure, et les règles de droit public.

De nombreux cas d'utilisation « indésirable » des deepfakes devraient en principe être couverts par la législation actuelle, notamment par le droit civil, comme les atteintes à la personnalité, et la production de deepfakes qui constitue en principe une violation des droits d'auteur et de la protection des données. D'autres utilisations de deepfakes relèvent du droit pénal, comme les atteintes à l'honneur, l'extorsion ou l'escroquerie ainsi que la pornographie, le harcèlement sexuel ou la pédopornographie. La manière dont le nouveau délit d'usurpation d'identité s'appliquera aux deepfakes demeure incertaine. Il existe en outre des zones grises, notamment en ce qui concerne les hypertrucages « embarrassants » ou « osés » qui pourraient être qualifiés pénalement comme de la pornographie ou du harcèlement. Et il faut souligner que, précisément dans le cas de recours aux deepfakes pour la violence sexualisée basée sur l'image, le tort causé par l'atteinte à l'honneur en combinaison avec l'autodétermination sexuelle n'est pas couvert par le droit pénal en vigueur.

Les plus grands défis résident dans l'application des droits civils et pénaux. Dans le monde d'Internet, les individus qui produisent des contenus sont souvent inconnus ou basés à l'étranger. Toute action à l'encontre des plateformes comporte en outre des incertitudes procédurales. De plus, compte tenu de leur diffusion rapide et potentiellement massive sur Internet, il est presque impossible de stopper la circulation des deepfakes par des moyens juridiques. Enfin, en particulier dans les procédures civiles, les frais de procédure peuvent avoir un effet dissuasif sur les victimes d'hypertrucages.

Les deepfakes jouent aussi un rôle dans le quotidien juridique et servent parfois de preuves dans les procédures judiciaires. La présentation de documents falsifiés devant les tribunaux n'est pas un phénomène nouveau en soi, des programmes comme Photoshop permettant déjà depuis longtemps de manipuler des textes ou des photos. Si la technologie deepfake est une nouvelle variante technique qui facilite la falsification vidéo ou audio, les questions qu'elle soulève n'ont rien de vraiment nouveau. En la matière, les dispositions juridiques existantes concernant le traitement des preuves falsifiées sont donc applicables.

Le recours à la technologie deepfake est également envisageable dans le cadre d'enquêtes des autorités judiciaires. Elle permet par exemple de reconstituer virtuellement une scène de crime ou le déroulement d'un délit à partir de vidéos de téléphones portables, de caméras de surveillance et de scanners corporels. Cependant, la production de pédopornographie par ordinateur pour accéder à des

forums en ligne qui exigent le téléchargement d'images ou de vidéos personnelles dans le cadre d'opérations d'infiltration est controversée. Cette question est particulièrement problématique parce que la législation interdit aux autorités judiciaires de commettre des infractions, même pour enquêter sur des délits (la pédopornographie produite à l'aide de deepfakes étant également, selon nous, couverte par la définition de la pédopornographie interdite). Quant aux cas cités précédemment, ils comportent des incertitudes procédurales et posent de nouvelles questions, notamment en ce qui concerne la manière de garantir le droit de participation à tous les stades de la procédure, y compris celui de « commettre virtuellement un crime », la manière de garantir le caractère vérifiable des preuves ainsi recueillies ou la manière dont et la forme sous laquelle les preuves numériques seront conservées dans les dossiers à l'avenir.

Enfin, une dernière question se pose en matière de réglementation des deepfakes : certains États comme les États-Unis ont adopté des approches réglementaires spécifiques à cette technologie pour régir les deepfakes (par ex. par le biais d'une obligation de déclaration). Il existe aussi d'autres approches réglementaires spécifiques (par ex. pour lutter contre la désinformation, y compris par le biais de deepfakes) ainsi que des approches de corégulation ou d'autorégulation totale des plateformes.

## Les deepfakes et le journalisme

La responsabilité des journalistes est un élément important qui revient régulièrement dans le débat. Selon l'analyse de nos interviews sur les deepfakes dans le journalisme, les rédactions des médias suisses étaient encore plutôt rarement confrontées à ce phénomène en automne/hiver 2022. La rareté relative des deepfakes dans le contexte suisse influence la manière dont ils sont présentés et abordés dans les établissements de formation et celle dont les représentants des institutions médiatiques eux-mêmes les perçoivent. Cependant, avec la diffusion et l'accessibilité croissantes des technologies de création deepfake, il semble réaliste de s'attendre à une multiplication des cas.

Dans la pratique journalistique, les deepfakes sont considérés comme un cas particulier technique de désinformation qu'il convient d'identifier dans le cadre de l'examen des faits. Du point de vue des personnes interrogées, les normes journalistiques de base sont une bonne référence pour gérer les deepfakes. Ces standards devraient être enseignés dans le cadre de la formation en journalisme, puis rappelés et appliqués dans la pratique. Cela implique également que les in-

formations provenant des réseaux sociaux doivent toujours être recoupées avec d'autres sources. Tous les médias étudiés ne disposent pas encore de personnel ou de services spécialisés pour traiter les cas complexes de vérification des vidéos. Mais certaines rédactions disposent déjà d'équipes ou de personnes qui vérifient les contenus des réseaux sociaux ou des vidéos (expertes ou experts en intelligence open source, fact checkers), et d'autres sont en train de mettre en place ce type de poste. Malgré tout, les ressources financières limitées dans le monde du journalisme rendent peu réaliste la création ou l'expansion d'équipes internes de vérification dans tous les médias. On peut également douter de la fiabilité des outils techniques de détection d'hypertrucages à long terme.

Dans l'ensemble, si les deepfakes sont majoritairement perçus comme un risque par les personnes interrogées, on constate que dans la pratique journalistique ils sont assimilés au problème plus large de la désinformation. Les personnes interrogées craignent que les hypertrucages ne renforcent encore la perte de confiance dans les médias et la remise en question infondée des faits. Elles considèrent que, en plus de l'examen des faits et des procédés de vérification eux-mêmes, la sensibilisation du public aux deepfakes et l'information sur ces procédés de contrôle sont des moyens constructifs de promouvoir la confiance dans les médias. Laisser aux seuls médias le soin de vérifier les informations et de prendre conscience de leur caractère manipulé ne semble donc pas suffire. Il faut plutôt sensibiliser l'ensemble de la société à cette manipulation et à la nécessité de remettre en question les sources, ce qui souligne l'importance de promouvoir l'éducation aux médias.

La formation journalistique présente aussi les deepfakes comme un élément de la gestion de la désinformation. En la matière, ses objectifs sont essentiellement la transmission aux futurs journalistes du savoir nécessaire pour vérifier l'information selon les normes de la profession et l'enseignement des techniques de manipulation du matériel source existantes. Les deepfakes servent dès lors aujourd'hui majoritairement d'exemples dans cette formation. Compte tenu de la difficulté à les détecter, la référence aux normes journalistiques de base prend encore plus d'importance. Ces standards fondamentaux sont enseignés lors de la formation, en association avec les deepfakes. Avec les progrès de la technologie et son accessibilité croissante, on peut s'attendre à ce que les hypertrucages prennent une place significative dans l'enseignement en tant que matière et compétence requise pour vérifier les informations complexes. Il paraît également incontournable que les professionnels des médias suivent des formations continues en matière d'évolution technologique et de moyens de vérification.

Jusqu'à présent, les rédactions suisses n'ont pas encore été victimes de deepfakes. Des stratégies de gestion des attaques ont déjà été discutées et mises en place pour répondre à la désinformation. Si les deepfakes sont considérés comme un risque supplémentaire, ils s'inscrivent dans le contexte de ce problème fondamental et de la méfiance envers le journalisme. Publiant davantage de matériel visuel, les journalistes qui travaillent dans le domaine de la télévision et de la vidéo courent un risque plus élevé d'être victime de deepfakes.

## Les deepfakes et la politique

La question des éventuelles implications politiques des deepfakes fait l'objet de vives discussions. Il existe des craintes persistantes à l'égard d'une utilisation abusive des deepfakes dans le but de nuire à la démocratie et à l'économie. C'est pourquoi nous avons étudié quel rôle les deepfakes sont susceptibles de jouer dans la politique suisse, puis dans l'économie.

À ce jour, la littérature scientifique sur l'utilisation des deepfakes en politique ne couvre généralement que des aspects partiels et n'aborde pas du tout la situation spécifique de la Suisse. Nous avons par conséquent d'abord mené une enquête auprès des parlementaires suisses et des membres de l'administration fédérale. Des scénarios détaillés sur le rôle potentiel des deepfakes dans la politique suisse ont ensuite été élaborés à partir de la revue de la littérature et des résultats de l'enquête. Enfin, chaque scénario a été associé à des recommandations de mesures de protection et de limitation des dommages.

La majorité des personnes interrogées lors de notre enquête dans les milieux officiels estiment que les deepfakes sont déjà présents dans les débats politiques, ou souhaitent que l'on accorde plus d'attention à ce sujet. En ce qui concerne les opportunités et les risques des deepfakes, leur position est claire : les personnes interrogées n'y voient quasiment que des risques. Leurs inquiétudes portent notamment sur les conséquences pour la démocratie suisse et les institutions politiques. Sur ces deux aspects, les personnes interrogées considèrent aussi majoritairement que la probabilité d'occurrence est élevée ou très élevée. Au nombre des dangers pertinents figurent le risque d'en être l'objet à titre personnel, le risque de tromperie et le risque qu'ils présentent pour les relations internationales. La majorité des participantes et participants estiment toutefois que la probabilité d'occurrence de tels risques est plutôt faible, voire très faible. Enfin, l'enquête montre également que les mesures de protection concrètes contre les deepfakes sont encore rares à l'heure actuelle.

Nous avons élaboré différents scénarios pour illustrer l'utilisation de deepfakes dans la politique suisse. En faisant des déclarations systématiques sur les agresseurs ou auteurs potentiels, le type d'attaque ainsi que le cercle des destinataires, ces scénarios révèlent que la plupart des formes d'agression qu'ils illustrent sont déjà réalisables à l'aide des technologies deepfake disponibles aujourd'hui. Il s'agit notamment du chantage, de l'intimidation et de l'atteinte à la réputation de responsables politiques, de fonctionnaires, etc. Les deepfakes peuvent également inciter à la haine et à la violence, porter atteinte à la réputation des institutions politiques et faciliter l'accès à des informations confidentielles en trompant les employées et employés ou même les systèmes d'authentification biométriques. Enfin, les scénarios explorent les effets potentiels des hypertrucages au niveau sociétal. Ils examinent aussi, en plus des moyens d'exercer une influence sur les élections et les processus décisionnels politiques, les moyens d'exacerber les tensions sociales, de nuire à la démocratie, de menacer la sécurité publique et d'influencer les relations internationales.

## Les deepfakes et l'économie

La criminalité économique s'appuie de plus en plus souvent sur les deepfakes, notamment pour dérober et usurper des identités. La multiplication de ce type de cas nous a donné l'occasion d'étudier ces cas de figure en Suisse. Dans ce chapitre, nous avons d'abord examiné le potentiel de risque et d'opportunité des deepfakes dans l'économie, puis identifié les mesures envisageables de protection et de limitation des dommages.

Comme l'indique la revue des médias et de la littérature, l'efficacité des deepfakes dans le domaine de la criminalité économique n'est plus à prouver. Au vu de leurs applications (technologiques) potentielles et du mode opératoire des organisations criminelles, il est raisonnable de supposer que les hypertrucages deviendront de plus en plus courants parmi les tactiques des criminels professionnels. Nous avons également élaboré des scénarios dans ce domaine visant à rendre ce phénomène concret. Ces scénarios illustrent les applications potentielles des deepfakes en matière de criminalité économique, notamment pour la fraude et l'usurpation d'identité, l'atteinte à la réputation des entreprises, la fraude publicitaire en ligne au moyen de profils synthétiques, jusqu'à la manipulation du marché et la manipulation des processus de décision démocratiques au détriment de l'économie ou de certains secteurs économiques. Les deepfakes permettent aussi de tromper les employées et employés et de déjouer les

systèmes d'authentification biométriques, de sorte que les agresseurs accèdent à des systèmes sécurisés pour, par exemple, introduire des rançongiciels à des fins de chantage ou pratiquer l'espionnage économique et industriel.

Une autre analyse du contenu des médias a révélé qu'au-delà du débat sur les risques, les deepfakes présentent également des possibilités d'application utiles dans différents domaines, essentiellement en matière de divertissement, de formation et de publicité. Dans ces secteurs, le recours aux technologies deepfake renouvelle l'industrie du divertissement, renforce la culture des fans, améliore la fidélisation de la clientèle, offre une nouvelle créativité satirique et entraîne des économies de coûts et de temps. Loin des possibilités d'exploitation à des fins économiques, de telles applications sont susceptibles de protéger les minorités et les identités, d'améliorer les offres de formation et de contribuer à accroître la motivation à apprendre et à renforcer les compétences médiatiques. Ces différentes opportunités s'accompagnent néanmoins aussi de nouveaux défis comme les questions de droits d'auteur.

## Conclusion et recommandations

Les nouvelles technologies de synthèse et de manipulation des médias occuperont à l'avenir une place de choix dans la vie quotidienne. Dans cette étude, nous avons examiné comment gérer cette « réalité médiatique modifiée » dans laquelle des contrefaçons trompeusement réalistes rivalisent avec des contenus médiatiques authentiques pour obtenir les faveurs du public. Tout d'abord, notre étude montre que la majorité de la population suisse considère actuellement les deepfakes comme un risque dans le domaine du journalisme et de la politique – et que, en particulier dans ces deux secteurs, ils sont aussi souvent assimilés à de la désinformation. Mais selon nos considérations sur le droit et l'économie, les hypertrucages peuvent être plus que cela. Si certaines utilisations des technologies deepfake violent effectivement la loi, dans d'autres cas, les hypertrucages sont protégés par la liberté d'expression, la liberté d'information, la liberté artistique ou le droit d'auteur. Dans le domaine de l'économie en particulier, il est clair que les technologies deepfake présentent aussi un énorme potentiel créatif et économique. L'analyse de la perception du public met en évidence à quel point la terminologie choisie est déterminante : lorsqu'il est question de médias synthétiques, les personnes interrogées expriment une opinion plus favorable sur l'objet de l'enquête que lorsqu'il est question de deepfakes.

Les effets indésirables des deepfakes ne peuvent être évités ou endigués par des mesures isolées d'ordre réglementaire ou technique. Nous proposons plutôt de combattre les dommages causés par les hypertrucages et de les prévenir par une combinaison de différentes mesures à plusieurs échelons et par différents acteurs. Ces mesures sont requises au niveau politique, mais aussi dans pratiquement tous les secteurs, car les deepfakes affecteront très probablement tout type d'organisation. Les exploitants de plateformes, les institutions médiatiques et les agences de presse ainsi que les acteurs du secteur de la communication seront également sollicités à cet égard. Enfin, les établissements de formation et la société dans son ensemble ont un rôle à jouer, car les citoyennes et citoyens capables de juger le contenu des médias avec une saine distance critique sont moins vite induits en erreur.

En revanche, nous exprimons une certaine réserve à l'égard des attentes suscitées par les progrès en matière de détection technique des deepfakes : cet espoir existe depuis de nombreuses années mais, à notre connaissance, aucun logiciel de détection n'est capable de concrétiser ces promesses à ce jour. Cela ne remet pas en cause le fait que la mise au point de méthodes de détection, tout comme celle de nouvelles approches d'authentification et de marquage, peut être intégrée dans une stratégie globale de gestion des deepfakes.

Dans notre étude, nous avons examiné une large gamme d'utilisations abusives des hypertrucages, et de solutions possibles pour y faire face. Sur cette base, nous préconisons *les mesures suivantes* :

**Réglementation des plateformes :** L'une des principales recommandations de l'étude est que les États poursuivent leurs efforts en matière de réglementation des plateformes. À cet égard, il ne faut pas se limiter à imposer l'obligation de supprimer ou de bloquer les deepfakes signalés en cas de soupçon fondé d'infraction aux exploitants de plateformes. Introduire une obligation supplémentaire consistant à mettre en place un système de notification des contenus illicites assorti de directives de transparence et de mécanismes de recours serait un pas essentiel pour renforcer les droits des victimes de deepfakes et des individus affectés par des suppressions non autorisées.

**Formation et responsabilité personnelle des citoyennes et citoyens :** Une autre recommandation importante consiste à renforcer les mesures éducatives et à mettre l'accent sur la responsabilité personnelle. En effet, même si les deepfakes n'ont pas encore eu l'effet tant redouté d'une perte de confiance généralisée dans les contenus médiatiques, les deepfakes et l'IA générative rendront de plus en plus difficile la différenciation entre contenu authentique et trucages.

L'éducation aux médias est donc essentielle pour permettre à chacune et chacun de continuer à faire cette distinction ou d'apprendre d'autres méthodes de vérification de la crédibilité d'un contenu, comme l'examen de la source. Pour que les citoyennes et citoyens puissent assumer leur responsabilité personnelle, des mesures d'accompagnement sont toutefois nécessaires, notamment dans le domaine de la formation et de la promotion des compétences médiatiques. Renforcer le financement public des services d'aide aux victimes spécialisés dans la cybercriminalité peut également y contribuer de manière importante, car cela permettra aux victimes, notamment en cas de cyberintimidation par la pornographie deepfake, de bénéficier d'un soutien plus efficace.

**Préparation aux deepfakes :** Dans tous les domaines de la société et tous les secteurs économiques, les organisations devraient se préparer à la multiplication des deepfakes. De nombreux acteurs s'intéressent déjà aux opportunités qu'offrent les technologies basées sur l'IA. En effet, le recours aux technologies deepfake est susceptible de contribuer à la réduction des coûts et de générer d'autres avantages économiques dans différents domaines. Mais les hypertrucages peuvent également être exploités à l'encontre de n'importe quelle organisation. Pour se prémunir contre cette cybercriminalité, toute organisation devrait procéder à une évaluation interne des risques et déterminer, le cas échéant, les mesures préventives et réactives à prendre. Parmi celles-ci figurent la formation continue, la création d'équipes spécialisées dans la réponse aux crises et le recours à des mécanismes d'authentification avancés.

**Journalisme :** Enfin, la contribution du journalisme à un avenir caractérisé par un usage responsable des technologies deepfake compte pour beaucoup : maintenir des normes journalistiques élevées permet à la fois de détecter les deepfakes et d'informer directement le public de contenus falsifiés. Cela renforce en outre la fonction sociale du journalisme grâce aux contenus produits selon des normes professionnelles et sur la base d'informations vérifiées.

Cependant, compte tenu de l'extrême rapidité de l'évolution technologique, il faut se préparer à ce que les propositions formulées ici soient bientôt dépassées. Si les milieux politiques et les autres secteurs auxquels notre étude s'adresse gardent à l'esprit cette dynamique, ils seront prêts à repérer les nouveaux développements et à y répondre par des actions et d'autres mesures adéquates.

# Sintesi

I deepfake sono contenuti audiovisivi (foto, video e audio) sintetizzati o manipolati con tecniche di intelligenza artificiale (IA), che sembrano autentici e nei quali in genere una persona dice o fa qualcosa che in realtà non ha mai detto o fatto. Il termine «deepfake», usato per designare contenuti mediatici sintetici e manipolati per la prima volta nel 2017, occupa ormai uno spazio fisso nel dibattito politico e mediatico.

Il bilancio che si può trarre, trascorsi alcuni anni dalla loro comparsa, è in chiaroscuro: vari deepfake politici e truffe basate su deepfake, come recentemente le chiamate shock, sembrano confermare alcuni timori. D'altro canto è sorta anche una serie di applicazioni utili basate su contenuti mediatici sintetici e manipolati. Per ora il temuto uso massiccio di deepfake in campagne di disinformazione o addirittura la grande apocalisse dell'informazione non si sono verificati.

Il nostro studio mira a valutare le opportunità e i rischi dei deepfake per la Svizzera. Di fronte ai progressi delle tecnologie deepfake viene spontaneo chiedersi anzitutto qual è lo *stato attuale della tecnologia*? *Cosa sanno i cittadini dei deepfake* e come valutano le opportunità e i rischi? Finora su questi punti non sono praticamente mai state condotte ricerche. Lo studio esamina inoltre come inquadrare l'uso dei deepfake *dal punto di vista giuridico*: cosa è consentito e dove stanno i limiti? E la legislazione vigente è sufficiente ad esempio per proteggere la popolazione contro l'uso improprio di deepfake? Tra gli elementi importanti nella gestione dei deepfake è menzionato regolarmente l'operato responsabile del *giornalismo*. Ma come sono percepiti i deepfake nel giornalismo svizzero e come gestiscono i deepfake i giornalisti? Suscita ampie discussioni anche la questione delle possibili *implicazioni politiche* dei deepfake: concretamente, che ruolo potrebbero svolgere i deepfake nella politica e di quali opzioni di azione dispongono la politica e gli altri attori per evitare effetti indesiderati? Recentemente viene attribuito un grande potenziale soprattutto all'uso dei deepfake per *scopi economici*: in questo contesto esaminiamo quindi sia le opportunità economiche che possono offrire le tecnologie deepfake sia gli scopi economici abusivi per cui possono essere impiegate e le (possibili) misure di protezione preventive e reattive. Oltre a offrire orientamenti sulle opportunità e sui rischi delle tecnologie deepfake, lo studio si prefigge anche di *illustrare possibilità di azione* per garantire una gestione responsabile dei deepfake, in modo tale da non intaccare la percezione delle opportunità e al tempo stesso perlomeno ridurre, se non evita-

re, i rischi. Lo studio cerca risposte a questi interrogativi in singoli capitoli di approfondimento, i cui principali risultati sono descritti brevemente qui di seguito:

- analisi della situazione e delle tendenze nell'ambito delle tecnologie deepfake,
- percezione dei deepfake nella popolazione,
- deepfake nel diritto,
- deepfake nel giornalismo,
- deepfake nella politica,
- deepfake nell'economia,
- raccomandazioni.

## Analisi della situazione e delle tendenze nell'ambito delle tecnologie deepfake

L'analisi delle possibilità tecniche attuali e prevedibili per creare deepfake evidenzia che, negli scorsi anni, sono stati fatti progressi considerevoli nell'ambito della sintesi e della manipolazione di contenuti *audiovisivi* (immagini, audio e testi) sfruttando l'IA.

Nell'ambito dei deepfake visivi, attualmente sono disponibili cinque tecniche:

- *facial reenactment*: manipolazione dell'espressione del volto;
- *face morphing*: fusione di più volti;
- *face swapping*: trasposizione o sostituzione di un volto con un altro;
- generazione facciale: creazione di volti che in realtà non esistono;
- *full body puppetry*: cambiamento della posa o dei movimenti di una parte del corpo o dell'intero corpo, come se fosse un burattino.

Nell'ambito dei deepfake *audio*, negli scorsi anni l'evoluzione è andata in direzione di risultati da un lato sempre più credibili, che oltretutto dall'altro possono essere raggiunti con sempre meno materiale di addestramento.

Già oggi è così possibile creare un'ampia gamma di contenuti deepfake. Le moderne tecnologie deepfake consentono di copiare il modo di parlare, l'espressione del volto e addirittura i movimenti del corpo di una persona – sempre a patto che i creatori dispongano di sufficienti dati di input (p. es. campioni della voce o fotografie del volto) e potenza di calcolo. Con i moderni generatori di testi è inoltre possibile imitare il modo di scrivere e la lingua di una persona, il che può essere utilizzato come input per i deepfake. È possibile sintetizzare da zero o modificare con una qualità sempre migliore anche sfondi e scenari. Oltre che deepfake che raffigurano esseri umani, ciò consente di creare anche scenari realistici, come il finto schianto di un aereo o una catastrofe naturale. Già oggi è possibile produrre con un onere relativamente esiguo deepfake audio di alta qualità. Generatori di immagini IA consentono di produrre anche immagini deepfake di persone note al generatore, in particolare di personalità pubbliche, immagini talmente realistiche da ingannare l'osservatore. Mediante l'addestramento dei generatori con ulteriore materiale o con l'aggiunta di proprie immagini associate a un comando di testo è possibile generare anche un numero a piacimento di altre persone.

Tentativi effettuati da noi stessi di creare video deepfake confermano però che, malgrado gli enormi progressi a livello di tecnologia, restano necessari un ampio know-how tecnico ed enormi capacità di calcolo nonché disponibilità di tempo e denaro per creare video deepfake credibili. Sottoscriviamo quindi le opinioni di chi parte dal presupposto che, attualmente e nel futuro prossimo, non vi sono segni del fatto che la massa della popolazione sia in grado di creare video deepfake realistici al punto da ingannare e possa «inondare» di tali video l'universo dei media. La situazione potrebbe mutare se gli attuali generatori di video deepfake diventassero più facili da utilizzare o se i metodi text-to-video venissero perfezionati e resi accessibili ad ampie fasce di popolazione. Tali metodi consentirebbero di creare video deepfake a piacimento mediante un comando di testo, analogamente a quanto fanno attualmente generatori di testi o immagini basati sull'IA, come ChatGPT o Midjourney. È presumibile che i più popolari programmi di questo genere adotteranno provvedimenti di protezione per evitare impieghi dannosi o eticamente inaccettabili. Tali provvedimenti di protezione possono però essere aggirati. È verosimile anche la nascita di altri generatori, che rinuncino a linee guida etiche. Benché il pericolo che i (social) media vengano «inondati» di video deepfake credibili non sia acuto, in futuro la loro creazione diventerà nettamente più semplice. Sulla scorta di questi risultati partiamo dal presupposto che, nel futuro prossimo, i video deepfake saranno prodotti soprattutto da singoli individui o da attori organizzati, che dispongono del know-how e delle capacità tecniche di calcolo necessari.

I capitoli seguenti mostrano tuttavia che, per produrre effetti dannosi, non è necessaria una grande quantità di deepfake. Un singolo deepfake può già causare ingenti danni di natura politica, giuridica, sociale o economica. E non è neanche necessaria l'intenzione di ingannare o danneggiare, come hanno dimostrato le reazioni a una fotografia manipolata del papa con indosso un piumino. La gente può essere fuorviata anche da contenuti mediatici creati a scopo d'intrattenimento.

Misure tecniche discusse quali rimedi comprendono metodi di autenticazione dei contenuti originali, di etichettatura dei contenuti deepfake nonché di riconoscimento dei video deepfake. Tutte queste misure presentano delle lacune. In linea di principio, sia l'autenticazione dei contenuti affidabili sia l'etichettatura dei contenuti deepfake possono essere aggirate tecnicamente. Oltre che causare ingenti danni, deepfake certificati erroneamente come affidabili potrebbero anche far vacillare la fiducia in tutti i documenti certificati. La possibilità di aggirare i provvedimenti tecnici di protezione sarebbe presumibilmente accessibile soprattutto agli attori (statali) che dispongono di particolarmente tante risorse. Al tempo stesso ciò potrebbe tuttavia significare che, in ogni caso, la massa dei deepfake in circolazione sarebbe etichettata e la massa dei contenuti affidabili sarebbe certificata e riconoscibile in quanto tale. Malgrado le sue lacune, in combinazione con altre misure la certificazione potrebbe quindi far parte di una strategia globale volta a contenere gli effetti indesiderati dei deepfake.

La nostra analisi della letteratura e i test che abbiamo effettuato mostrano che gli attuali rilevatori di deepfake accessibili liberamente non sono affidabili. Miglioramenti non sono in vista: si prevede infatti che i generatori saranno sempre un passo in avanti rispetto ai rilevatori. Visti gli investimenti nello sviluppo di rilevatori e le sfide tecniche con cui sono confrontate le reti *GAN* (Generative Adversarial Networks) e altre tecnologie deepfake, non è tuttavia escluso che, in futuro, la precisione dei metodi migliorerà.

## Percezione dei deepfake nella popolazione svizzera

Finora l'atteggiamento della popolazione nei confronti dei deepfake non è praticamente stato oggetto di ricerche. Con un'indagine rappresentativa abbiamo esplorato come sono percepiti i deepfake in seno alla popolazione svizzera. A tal fine abbiamo rilevato le esperienze degli svizzeri con i deepfake, la loro valutazione delle opportunità e dei rischi della tecnologia deepfake e la loro capacità di distinguere i deepfake dai video reali.

Nel settembre 2023, per molte persone il termine deepfake restava un'incognita. Solo poco più della metà degli intervistati ne aveva già sentito parlare. Quasi la metà degli intervistati sosteneva di aver già visto dei deepfake. Solo una piccolissima minoranza aveva già avuto esperienze di creazione (2 %) e diffusione di deepfake (3 %). Da un lato ciò significa che la percezione dell'argomento è dettata anzitutto da quanto riferiscono i media e non si basa su esperienze proprie. Ciò offre però anche la possibilità di educare a una gestione intelligente della tecnologia deepfake mediante offerte di formazione e informazione.

La popolazione svizzera associa i deepfake soprattutto ai rischi, praticamente indistintamente in tutti i gruppi sociodemografici. Svolge tuttavia un ruolo il presunto effetto dei deepfake sull'opinione di terzi: più si parte dal presupposto che gli altri siano influenzati dai deepfake, più si considera elevato il rischio dei deepfake per la società. Emerge quindi un'idea piuttosto vaga degli effetti dei deepfake sulla società e questa vaghezza aumenta la percezione del rischio. L'uso dell'espressione «media sintetici» anziché «deepfake» ha invece un effetto positivo sulla percezione delle opportunità della tecnologia deepfake. Nel complesso, i risultati indicano che in seno alla popolazione vi è ancora molta insicurezza attorno ai deepfake. Spicca inoltre il fatto che, rispetto agli uomini, le donne giudicano superiori i rischi individuali – ossia la messa in pericolo della sfera privata o il rischio di restare vittime di un deepfake.

I nostri esperimenti rivelano che, in Svizzera, per la gente è pressoché impossibile distinguere i deepfake dai video reali, in particolare se sono di qualità da buona a ottima. L'alfabetizzazione mediatica generale è quindi particolarmente importante. lo studio evidenzia che *l'alfabetizzazione ai social media*, ossia la competenza di muoversi tra i nuovi media in Internet, ha un effetto positivo sulla competenza di riconoscere i deepfake. L'alfabetizzazione mediatica generale e dimestichezza con Internet non hanno invece alcun effetto. Il nostro breve intervento di alfabetizzazione, ossia una spiegazione di come riconoscere i deepfake subito prima di mostrare i video, non ha avuto alcun effetto sulla capacità di riconoscerli. Non si verifica tuttavia neanche un *effetto «backfire»*. L'intervento di alfabetizzazione non induce quindi gli intervistati a diventare ipercritici e di conseguenza a tendere a prendere per deepfake anche i video reali.

## Deepfake nel diritto

Il diritto svolge un duplice ruolo in relazione ai deepfake: da un lato li protegge quali forme di espressione artistica o di libertà di opinione, mentre dall'altro ha

anche l'importante funzione di offrire protezione contro gli effetti dannosi e/o indesiderati dei deepfake, in particolare mediante gli strumenti del diritto penale, della tutela del diritto d'autore e della protezione della personalità e dei dati come pure le disposizioni (in materia di media) di diritto pubblico. Da un lato i deepfake e i loro autori possono così far valere una protezione della libertà di opinione, d'informazione e dell'arte, poiché fino a un certo punto anche le esternazioni non veritiere, false o fuorvianti beneficiano di una protezione fondamentale. In determinate circostanze i deepfake possono inoltre essere protetti dal diritto d'autore. Questa protezione non è però assoluta: i diritti fondamentali possono essere limitati a determinate condizioni, segnatamente per tutelare i diritti di terzi. Tra i punti focali dell'analisi giuridica nell'ambito del presente studio figura quindi la protezione giuridica contro gli usi o gli effetti dannosi dei deepfake.

In Svizzera, attualmente non esiste un disciplinamento specifico relativo ai deepfake, ma bisogna ottemperare a diverse prescrizioni generali a seconda dell'uso previsto. Il termine «deepfake» e altre designazioni utilizzate in questo contesto non sono infatti definiti nel diritto, ma rientrano in parte in nozioni esistenti. Occorre distinguere tra diritto civile e penale (e i relativi codici procedurali) nonché diritto pubblico.

Molte delle applicazioni «indesiderate» dei deepfake dovrebbero essere contemplate dal diritto vigente. Nel diritto civile sono possibili segnatamente lesioni della personalità, inoltre normalmente la creazione di deepfake rappresenta anche una violazione dei diritti garantiti dal diritto d'autore e dalla legislazione sulla protezione dei dati. Alcune applicazioni dei deepfake rientrano anche nel diritto penale: entrano in considerazione in particolare i delitti contro l'onore, l'estorsione o la truffa nonché la pornografia, le molestie sessuali o la pornografia infantile. Come sarà applicata la nuova fattispecie penale dell'usurpazione dell'identità nel caso dei deepfake non è ancora noto. Vi sono tuttavia zone grigie, segnatamente per quanto riguarda la qualificazione penale come pornografia o molestie dei deepfake «compromettenti» o «discinti». Emerge anche che, soprattutto in caso di uso dei deepfake quale violenza sessualizzata basata su immagini, il contenuto illegale consistente nella lesione dell'onore in combinazione con l'autodeterminazione sessuale non è contemplato dal diritto penale vigente.

Le maggiori sfide riguardano l'esercizio delle pretese civili e penali. In Internet, gli autori dei reati sono spesso sconosciuti o hanno sede all'estero. Anche le azioni contro le piattaforme sono irte di incertezze procedurali. Vista la diffusione rapida e potenzialmente virale dei contenuti su Internet, è quasi impossibile «fermare» con mezzi legali un deepfake una volta che è stato messo in circolazione. Infine, soprattutto nei procedimenti civili, i costi procedurali possono essere un deterrente per le vittime di deepfake.

I deepfake possono svolgere un ruolo anche nella quotidianità legale, ad esempio come prove nei procedimenti giudiziari. Di per sé la presentazione di documenti falsificati in tribunale non è una novità. Da molti anni, programmi come Photoshop consentono di manipolare testi o fotografie. La tecnologia deepfake è una nuova variante tecnica, che facilita la falsificazione di video e audio, ma non solleva interrogativi fondamentalmente nuovi. Sono pertanto applicabili le norme giuridiche vigenti in materia di prove falsificate.

È anche ipotizzabile che la tecnologia deepfake possa essere utilizzata dalle autorità di perseguimento penale per chiarire i reati. La tecnologia deepfake può ad esempio essere utilizzata per creare una scena del crimine virtuale o per ricostruire il corso degli eventi a partire da video di telefoni cellulari e telecamere di sorveglianza nonché da scansioni corporee. La possibilità che investigatori sotto copertura utilizzino materiale pedopornografico generato al computer per ottenere l'accesso a forum online che richiedono il caricamento di proprie immagini o video è oggetto di un dibattito controverso. Se quest'ultima possibilità è problematica soprattutto perché, secondo il diritto vigente, le autorità di perseguimento penale non sono autorizzate a commettere esse stesse reati neanche per indagare su reati (secondo noi, anche la pornografia infantile creata con deepfake rientra nella definizione di pornografia infantile vietata), gli altri impieghi ipotizzati sollevano incertezze procedurali: come garantire segnatamente i diritti di partecipazione in tutte le fasi del procedimento e quindi anche in caso di «ispezioni virtuali della scena del crimine», come verificare le prove raccolte in questo modo o come e in quale forma archiviare in futuro le prove raccolte digitalmente?

Infine, vi è la questione del disciplinamento dei deepfake: alcuni Paesi, come gli Stati Uniti, perseguono approcci normativi basati sulla tecnologia, che disciplinano in modo specifico i deepfake (p. es. tramite un obbligo di dichiarazione). Vi sono però anche approcci normativi tematici (p. es. volti a lottare contro la disinformazione – anche attraverso i deepfake) nonché approcci di coregolamentazione o di completa autoregolamentazione delle piattaforme.

## Deepfake nel giornalismo

Nella discussione sulla gestione dei deepfake è regolarmente menzionato come elemento importante l'operato dei giornalisti. L'analisi delle interviste sui deepfake nel giornalismo mostra che, nell'autunno/inverno 2022, le redazioni delle organizzazioni mediatiche svizzere erano ancora confrontate piuttosto raramente

con i deepfake. La rarità dei deepfake, soprattutto nel contesto svizzero, influenzava quindi il modo in cui erano affrontati nelle organizzazioni di formazione dei giornalisti e valutati dai rappresentanti delle organizzazioni mediatiche stesse. Con la crescente diffusione e accessibilità della tecnologia per creare deepfake sembra però assolutamente realistico che, in futuro, i casi saranno più frequenti.

Nella pratica giornalistica, i deepfake sono visti come un caso tecnico particolare di disinformazione, che deve essere riconosciuto nel contesto della verifica delle fonti (fact-checking). Stando agli intervistati, quando si ha a che fare con i deepfake è utile farsi guidare da norme e standard giornalistici di base, che dovrebbero essere appresi durante la formazione e ricordati e applicati nella pratica. Tra tali norme e standard figura anche il fatto che le informazioni provenienti dai social network dovrebbero sempre essere verificate in base ad altre fonti. Le organizzazioni mediatiche esaminate segnalano tuttavia la necessità di persone o dipartimenti specializzati che si occupino della verifica dei video nei casi complicati. Alcune redazioni dispongono già di tali team o persone per la verifica dei contenuti provenienti dai social network o dai video (esperti di intelligence open source, fact-checker) o stanno per creare tali posizioni. A causa delle scarse risorse finanziarie del giornalismo, la creazione o l'espansione di team di verifica interni non è probabilmente un'opzione realistica per tutte le organizzazioni mediatiche. È inoltre discutibile quanto saranno ancora affidabili in futuro gli strumenti tecnici di riconoscimento dei deepfake.

Nel complesso, gli intervistati percepiscono i deepfake perlopiù come un rischio. Nella pratica giornalistica non si fa tuttavia praticamente alcuna differenza rispetto al più ampio problema della disinformazione. Gli intervistati temono che i deepfake aumentino ulteriormente la perdita di fiducia nei media e la messa in discussione immotivata dei fatti da parte di una parte degli utenti. Oltre ai processi di verifica e al fact-checking, gli intervistati ritengono quindi che per promuovere la fiducia nei media in modo costruttivo occorra sensibilizzare il pubblico sui deepfake e informare sui processi di verifica. Non sembra quindi sufficiente che le informazioni siano verificate e che vi sia una consapevolezza delle informazioni manipolate solo da parte dei media: è necessaria una consapevolezza delle informazioni manipolate e della necessità di sottoporre le fonti a un esame critico in tutta la società, il che non fa che ribadire l'importanza di promuovere l'alfabetizzazione mediatica.

Anche la formazione giornalistica tratta i deepfake come parte della gestione della disinformazione. Gli obiettivi principali della formazione sono quelli di fornire ai futuri giornalisti la conoscenza di come verificare le informazioni secondo le norme professionali e di quali possibilità tecniche esistono per manipolare il

materiale di partenza. Attualmente, nei corsi i deepfake sono trattati prevalentemente come esempi. La complessità dell'identificazione dei deepfake fa sì che sia sempre più importante orientarsi alle norme e agli standard giornalistici di base. Queste norme sono apprese durante la formazione e associate ai deepfake. Sulla scia del miglioramento e della maggiore accessibilità della tecnologia è ipotizzabile che in futuro la tematica dei deepfake e le competenze necessarie per verificare informazioni complesse diventeranno ancora più importanti nella formazione. Inoltre sembra indispensabile una formazione continua dei professionisti dei media in relazione agli sviluppi tecnici e alle possibilità di verifica.

## Deepfake nella politica

Un argomento molto discusso è quello delle possibili implicazioni politiche dei deepfake. Vi sono timori persistenti sull'uso improprio dei deepfake per danneggiare la democrazia e l'economia. Per questo motivo abbiamo analizzato il possibile ruolo dei deepfake dapprima nella politica svizzera e poi nell'economia.

Nella nostra indagine tra i parlamentari svizzeri ed esponenti dell'Amministrazione federale, la maggioranza degli intervistati ha dichiarato che i deepfake sono già oggetto di discussione a livello politico o ha auspicato che si presti maggiore attenzione alla questione. Per quanto riguarda le opportunità e i rischi, è emerso un quadro chiaro: gli intervistati vedono quasi esclusivamente rischi. Vi è particolare preoccupazione per l'impatto sulla democrazia svizzera e sulle istituzioni politiche. Per entrambi gli aspetti, la maggior parte degli intervistati giudica inoltre alta o molto alta la probabilità che si verifichi tale impatto. Tra i principali rischi sono menzionati anche quello che circoli un deepfake su di sé, quello di essere ingannati tramite un deepfake nonché i rischi derivanti dai deepfake per le relazioni internazionali. La maggior parte delle risposte indica tuttavia che la probabilità che tali rischi si verifichino può essere classificata come bassa o molto bassa. L'indagine rappresentativa mostra infine anche che attualmente sono ancora adottate troppo raramente misure di protezione concrete contro i deepfake.

Per concretizzare il possibile uso di deepfake nella politica svizzera sono stati elaborati diversi scenari, che formulano osservazioni sistematiche sui possibili aggressori o autori, sul tipo di attacco e sui destinatari. Emerge che buona parte delle forme di attacco presentate negli scenari può già essere realizzata con le tecnologie deepfake disponibili oggi. Tra queste figurano l'estorsione o l'intimidazione e i danni alla reputazione di politici, funzionari, ecc. I deepfake possono anche contribuire a incitare all'odio e alla violenza. Sono adatti anche per

danneggiare la reputazione delle istituzioni politiche e per ottenere informazioni riservate ingannando i dipendenti ed eventualmente i sistemi di autenticazione biometrica. Gli scenari affrontano infine i possibili effetti a livello sociale. Oltre all'influsso sulle elezioni e sui processi decisionali politici, vengono menzionate anche le possibilità di esacerbare le tensioni sociali, danneggiare la democrazia, mettere a rischio la sicurezza pubblica e influenzare le relazioni internazionali.

## Deepfake nell'economia

I deepfake sono sempre più utilizzati per scopi di criminalità economica, ad esempio in relazione a frodi e furti di identità. L'aumento di questi casi ci ha spinti a indagare su queste costellazioni in Svizzera.

La panoramica dei media e della letteratura mostra che i deepfake vengono già utilizzati con successo per scopi di criminalità economica. Sulla base delle possibilità (tecnologiche) di applicazione dei deepfake e del modus operandi dei gruppi criminali, si può ipotizzare che i deepfake entreranno sempre più nel repertorio di attacco degli attori criminali professionisti. Anche in questo campo, per concretizzare la situazione sono stati elaborati scenari, che illustrano i potenziali usi dei deepfake nell'ambito della criminalità economica, in particolare per le frodi e i furti di identità, i danni alla reputazione delle aziende, le frodi pubblicitarie online mediante profili sintetici o ancora la manipolazione del mercato e dei processi decisionali democratici a scapito dell'economia o di determinati settori economici. I deepfake possono inoltre essere utilizzati per ingannare i dipendenti e aggirare i sistemi di autenticazione biometrica, consentendo agli aggressori di accedere a sistemi protetti e installarvi ad esempio ransomware a scopo di estorsione o di condurre spionaggio economico e industriale.

Un'analisi dei contenuti mediatici ha inoltre dimostrato che, al di là del dibattito sui rischi, i deepfake offrono anche opportunità di impiego intelligente in vari campi. Le opportunità s'iscrivono principalmente nei settori dell'intrattenimento, dell'istruzione e della pubblicità. Il ricorso alle tecnologie deepfake in questi settori può permettere di creare nuove offerte di intrattenimento, sostenere la cultura dei fan e fidelizzare i clienti, risparmiare tempo e denaro e creare nuove offerte satiriche. Oltre alle possibilità di sfruttamento economico, queste applicazioni possono anche consentire la protezione delle minoranze e delle identità, essere utilizzate per migliorare i programmi educativi e contribuire ad aumentare la motivazione all'apprendimento e a rafforzare l'alfabetizzazione mediatica. Le varie opportunità possono tuttavia comportare anche nuove sfide, come i problemi legati al diritto d'autore.

## Conclusioni e raccomandazioni

Le nuove tecnologie di sintesi e manipolazione dei media sono destinate a occupare un posto fisso nella cultura quotidiana. In questo studio abbiamo discusso di come affrontare in futuro questa «realtà mediatica mutata», in cui falsi realistici al punto da ingannare si contendono il favore dei destinatari con contenuti mediatici originali. In primo luogo, lo studio rivela che, attualmente, la maggioranza della popolazione svizzera, gli ambienti giornalistici e la politica considerano i deepfake un rischio. I deepfake sono spesso classificati come disinformazione, in particolare nel giornalismo e nella politica. Le nostre considerazioni sul diritto e sull'economia mostrano tuttavia che i deepfake possono essere più di questo: se alcuni usi delle tecnologie deepfake rientrano effettivamente nell'ambito dell'illecito, in altri casi i deepfake sono protetti dalla libertà di opinione, d'informazione e dell'arte o dal diritto d'autore. In particolare se si considerano i deepfake nell'economia, emerge chiaramente che le tecnologie deepfake hanno anche un enorme potenziale creativo ed economico. Uno sguardo alla percezione pubblica evidenzia quanto la terminologia scelta determini la direzione del discorso: quando si parla di media sintetici, gli intervistati valutano l'oggetto di indagine in modo più positivo rispetto a quando si parla di deepfake.

Le conseguenze indesiderate dei deepfake non possono essere evitate o contenute con singole misure normative o tecniche. Proponiamo invece che i danni causati da deepfake siano combattuti e prevenuti mediante una combinazione di diverse misure adottate a vari livelli e da diversi attori. È necessaria un'azione da parte della politica, ma anche delle organizzazioni di praticamente tutti i settori, essendo molto probabile che le conseguenze dei deepfake interessino tutte le organizzazioni. Sono chiamati a intervenire anche i gestori delle piattaforme, le organizzazioni dei media e le agenzie di stampa come pure gli attori dell'industria della comunicazione. Non da ultimo sono chiamate in causa le istituzioni educative e la società in generale: cittadini che valutano i contenuti dei media con una sana distanza critica sono infatti meno suscettibili di essere ingannati.

Siamo invece piuttosto scettici di fronte all'eccessiva fiducia nei progressi del riconoscimento tecnico dei deepfake: questa speranza esiste da molti anni, eppure non siamo a conoscenza di alcun software di rilevamento in grado di mantenere le promesse. Ciò non significa tuttavia negare che l'ulteriore sviluppo di metodi di rilevamento così come di nuovi metodi di autenticazione ed etichettatura possa essere un elemento di una strategia a tutto campo per affrontare i deepfake.

Nel nostro studio abbiamo analizzato un ampio ventaglio di possibili conseguenze abusive del deepfake ed elaborato raccomandazioni per affrontarle. Formuliamo pertanto le seguenti *raccomandazioni d'azione*.

**Disciplinamento delle piattaforme:** tra le principali raccomandazioni dello studio figura quella di portare avanti gli sforzi statali per disciplinare le piattaforme. In questo contesto è consigliabile non solo introdurre l'obbligo per i gestori delle piattaforme di cancellare o bloccare i deepfake segnalati in caso di sospetto motivato di violazione della legislazione. I gestori delle piattaforme dovrebbero anche istituire un sistema di segnalazione dei contenuti illeciti, che includa requisiti di trasparenza e possibilità di opposizione: ciò rafforzerebbe i diritti sia delle vittime dei deepfake sia di coloro che sono colpiti da cancellazioni ingiustificate.

**Educazione e responsabilità individuale dei cittadini:** un'altra raccomandazione importante è quella di rafforzare le misure educative e ribadire la responsabilità individuale dei cittadini. Anche se la temuta perdita generalizzata di fiducia nei contenuti mediatici come conseguenza dei deepfake non si è ancora concretizzata, con i deepfake e l'IA generativa sarà sempre più difficile distinguere i contenuti autentici dai falsi. Questo rende ancora più importante l'alfabetizzazione mediatica, in modo che, in futuro, ogni singolo individuo possa ancora fare questa distinzione o impari altri metodi per verificare la credibilità dei contenuti, come la verifica delle fonti. Affinché i cittadini possano assumersi le proprie responsabilità, sono tuttavia necessarie misure di accompagnamento, in particolare nel campo dell'educazione, per promuovere l'alfabetizzazione mediatica. Anche il rafforzamento dei finanziamenti statali destinati ai consultori per le vittime specializzati nella criminalità informatica può dare un contributo importante, consentendo ad esempio di offrire un sostegno più efficace alle vittime di cyberbullismo attraverso la pornografia deepfake.

**Preparazione ai deepfake:** le organizzazioni di tutti gli ambiti della società e di tutti i settori dell'economia dovrebbero prepararsi alla crescente ubiquità dei deepfake. Molti attori stanno già lavorando sulle opportunità offerte dalla tecnologia basata sull'IA. In vari settori, l'uso delle tecnologie deepfake può contribuire a ridurre i costi e a produrre altri vantaggi economici. D'altro canto, i deepfake possono anche essere usati in modo improprio contro qualsiasi organizzazione. Per difendersi dall'uso dei deepfake da parte dei criminali informatici, le organizzazioni dovrebbero effettuare una valutazione interna dei rischi e, se necessario, individuare le misure preventive e reattive da adottare. Queste possono includere misure di formazione continua, la creazione di team specializzati nella risposta alle crisi o l'uso di misure di autenticazione avanzate.

**Rispetto degli standard giornalistici:** infine, il giornalismo può fornire un contributo prezioso a un futuro contraddistinto da un uso responsabile delle tecnologie deepfake: il rispetto degli standard giornalistici può servire sia a riconoscere i deepfake sia a informare direttamente la popolazione sui contenuti falsificati. Consente inoltre al giornalismo di svolgere meglio la sua funzione sociale di fornire contenuti prodotti secondo standard professionali sulla base di informazioni verificate.

Di fronte all'enorme velocità del cambiamento tecnologico è però prevedibile che le proposte avanzate qui saranno presto superate dalla realtà. I politici e gli altri attori a cui ci rivolgiamo dovrebbero essere consapevoli di questa dinamica e pronti a riconoscere i nuovi sviluppi e a reagire con ulteriori iniziative e misure appropriate.

# 1. Einleitung und Kontext

*Murat Karaboga*

## 1.1. Hintergrund und Zielsetzung der Studie

Bei gefälschten und manipulierten Inhalten handelt es sich mitnichten um ein neues Thema: Die Geschichte von Fälschungen reicht Jahrtausende zurück. Schon in der Antike wurden Münzen, Schriftrollen und Schmuck gefälscht. Mit dem verstärkten Aufkommen von Textdokumenten ab dem Mittelalter begann auch die Fälschung von Urkunden, Briefen, anderen Rechtstexten. Überall dort, wo das Fälschen und Manipulieren einen Vorteil versprach, wurde es durchgeführt. Mit dem Aufkommen der Fotografie seit den 1800er-Jahren wurden Fälschungen allmählich um manipulierte Fotografien ergänzt. Mit der zunehmenden Popularität und Verbreitung von Bewegtbildmedien kamen im 20. Jahrhundert Möglichkeiten der Videomanipulation hinzu. Schliesslich wurde mit der Digitalisierung Bildbearbeitung, die im analogen Zeitalter noch äusserst ressourcenaufwendig war, für die breite Masse der Computernutzenden zugänglich. Sinnbildlich dafür steht die Manipulation von Bildern mit der Software Photoshop, das sog. *Photoshoppen*, das sich über die Jahre zu einem integralen Bestandteil der digitalen Kultur entwickelte (Farid/Schindler 2020; Paris/Donovan 2019).

Seitdem im Herbst 2017 auf Reddit pornografische Videos hochgeladen wurden, in denen die Gesichter der Pornodarstellerinnen mit den Gesichtern prominenter Frauen ausgetauscht waren (sog. «deepnudes»), hat sich eine Debatte um diese neue Form der Medienmanipulation entfaltet.

Dieses unter Rückgriff auf den Namen des Users («deepfakes»), der die Inhalte hochgeladen hatte, fortan als Deepfake (teilweise auch als «Deep Fake») bezeichnete Phänomen fand schon bald zahlreiche Nachahmer im Internet. Auch andere User begannen, die Gesichter von Pornodarstellerinnen durch das Gesicht nicht nur von prominenten Persönlichkeiten, sondern auch durch die Gesichter beliebiger anderer Frauen zu ersetzen. Das Überraschende oder Schockierende daran war nicht die prinzipielle Möglichkeit der Videofälschung oder die übergriffige, sexuelle Stigmatisierung von Frauen. Neu war, dass die Fälschung von Bewegtbildern, für die es zuvor – bspw. in Form von *Computer Generated Imagery* (CGI) – noch enorm aufwendiger Rechenkapazitäten bedurfte, nunmehr vom heimischen Computer aus möglich war. Dies wurde durch eine neue KI-basierte Technologie namens *Generative Adverserial Networks* (GAN) ermöglicht, die zwei künstliche neuronale Netzwerke konkurrieren lässt, um glaubwürdige Bewegtbildinhalte zu generieren.

Obwohl die auf Reddit geteilten pornografischen Deepfakes eine niedrige Auflösung und Qualität hatten und daher häufig bereits mit blossem Auge als Fake zu erkennen waren, entwickelte sich eine Debatte um die Zukunft der Glaubwürdigkeit von bildbasierten Medien im Allgemeinen. Gegenstand dieser neuen wissenschaftlichen und gesellschaftlichen Debatte war, dass Fortschritte in der künstlichen Intelligenz (KI) schon bald die Erstellung von Deepfake-Videos ermöglichen würden, die verschiedenste Menschen in Situationen abbilden, die so zwar nie stattgefunden haben, aber nicht mehr als Fälschung zu erkennen wären. Die einfache Zugänglichkeit der Deepfake-Anwendungen würde schliesslich zu einer nie dagewesenen Flut an gefälschten und manipulierten Videos führen.

Nachdem die Debatte anfangs noch auf die Kritik an und die Furcht vor Deepfake-Pornografie sowie misogyne Verwendungen fokussierte, kamen in den Folgejahren weitere Bedenken hinzu. Ein wichtiger Aspekt war die missbräuchliche Nutzung von Deepfakes für Desinformation zur Schädigung der Demokratie, insbesondere durch die Manipulation von Wahlen. Einen bedeutenden Anteil (vgl. auch Abbildung 1) an der Popularität der Thematik hatte ein Deepfake-Video, in dem der US-Schauspieler und Regisseur Jordan Peele den früheren US-Präsidenten Barack Obama mimt, wie er über die Gefahren von Desinformation spricht. Durch die Imitation der Mikromimik, Gestik, der Stimme und Sprechweise Obamas, war es Peele gelungen, ein täuschend echtes Deepfake-Video zu erschaffen, das vielen Menschen demonstrierte, wozu einmal Deepfakes – auch ohne die Einbindung eines professionellen Schauspielers – in der Lage sein könnten (BuzzFeedVideo 2018). Im Laufe der Jahre wurden viele weitere Deepfake-Videos mit Aufklärungsabsicht veröffentlicht, die z.B. die frühere britische Königin Elizabeth II. und Mark Zuckerberg zeigten (Rahim 2020).

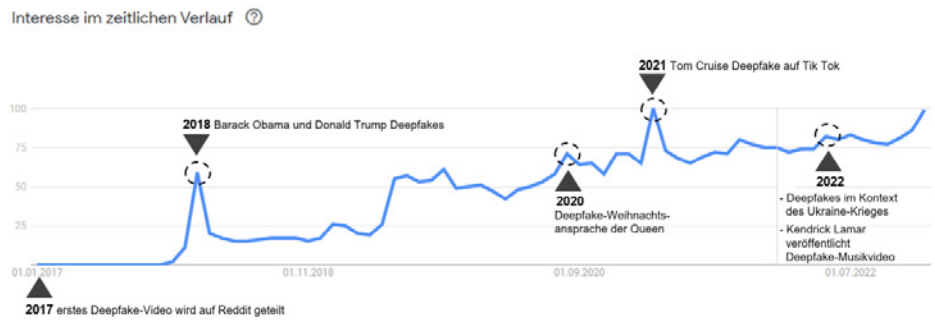


**Abbildung 1:** Google-Suchanfragen «Deepfake» weltweit und Fälle medienwirksamer Deepfakes (Google Trends; 01.01.2017–01.11.2023)

Als das Thema Deepfakes von der Forschung aufgegriffen wurde, entstand eine Vielzahl an Publikationen, die weitere Missbrauchsmöglichkeiten identifizierten, darunter die Destabilisierung politischer Prozesse, die Befeuerung inner- oder zwischenstaatlicher Konflikte, die Diskreditierung demokratischer Institutionen oder von Politikerinnen und Politikern. In einer Zukunft, in der Menschen synthetische und manipulierte Inhalte nicht mehr von originalen Inhalten unterscheiden können, so die vielleicht weitreichendste Befürchtung, könne es gar zu einer Erosion des Vertrauens in bestehende gesellschaftliche Ordnungen und Strukturen kommen, also eine gesellschaftliche Situation, in der von vornherein allen medialen Inhalten misstraut wird und ein Klima des Misstrauens nicht nur in Medieninhalte, sondern auch gegenüber Journalistinnen und Journalisten, politischen Institutionen und letztlich der Demokratie selbst vorherrschend sein wird.

Derweil wurde die öffentliche Debatte von technologischen Entwicklungen weiter vorangetrieben. Fortschritte im Voice Cloning führten zur Veröffentlichung von Anwendungen zur Erstellung von Deepfake-Audios, in denen die Stimme einer Zielperson auf glaubwürdige Weise imitiert oder verändert wird. Damit wurde nicht nur die Imitation der Stimme einer Zielperson in einem Deepfake-Video möglich, sondern auch die Umgehung der biometrischen Spracherkennung bei Banken und die Steigerung der Glaubwürdigkeit des Enkeltricks und von Schockanrufen. Des Weiteren wurden seit Ende 2022 mehrere KI-basierte Bildgeneratoren wie Midjourney und Dall-E veröffentlicht, die demonstrierten, dass inzwischen die Erstellung von täuschend echten synthetischen Bildern mittels Textbefehlen möglich ist. Aufsehenerregend waren synthetische Bilder, die den Papst in Daunenjacke und Donald Trump im Gerangel mit der Polizei zeigen. Ebenfalls Ende 2022 wurde ein auf der *generative pre-trained transformer*-Technologie basierender Chatbot namens ChatGPT veröffentlicht. Unter der Vielzahl der Nutzungsmöglichkeiten des Chatbots zeigte sich, dass mithilfe von ChatGPT auch die Imitation der Sprech- und Schreibweise eines Menschen möglich ist.

Seit dem Auftreten des ersten Deepfakes lässt sich eine gemischte Bilanz ziehen: Mehrere politische Deepfakes und Deepfake-basierte Betrugsfälle scheinen manche der über die Jahre geäusserten Befürchtungen zu bestätigen. Gleichzeitig sind auch eine Reihe von nützlichen Anwendungen, insbesondere im Unterhaltungsbereich, entstanden, die auf synthetischen und manipulierten Text-, Bild- und Audioinhalten basieren. Die befürchtete Zunahme von Deepfakes in Desinformationskampagnen oder gar die grosse Informationsapokalypse sind jedenfalls bislang ausgeblieben. Doch wie ist der gegenwärtige Stand bei Deepfake-Technologien? Wie nehmen Menschen in der Schweiz Deepfakes

wahr? Inwiefern ist der Journalismus in der Schweiz von Deepfakes betroffen und wie gehen Journalistinnen und Journalisten mit den diskutierten Herausforderungen um? Welche Auswirkungen könnten Deepfakes auf die Politik und die Wirtschaft haben? Bringen Deepfakes auch Chancen mit sich? Wie können sich Betroffene rechtlich gegen Deepfakes wehren und inwiefern ist das Schweizer Recht generell gegen Deepfakes gewappnet? Und schliesslich die womöglich wichtigste Frage: Welche Handlungsmöglichkeiten haben Entscheidungstragende aus Politik und anderen Bereichen, um die mit Deepfakes zusammenhängenden Chancen und Risiken zu adressieren?

Das Ziel der vorliegenden Studie ist es, diese und weitere Fragen zu beantworten und damit fundiertes Orientierungswissen zum Umgang mit Deepfake-Technologien aus interdisziplinärer Perspektive und für verschiedene Akteure bereitzustellen, um den gesellschaftlichen und politischen Diskurs über die Chancen und Herausforderungen von Deepfake-Technologien zu befördern.

## 1.2. Zielsetzung

Die vorliegende Studie verfolgt das Ziel, fundierte Antworten auf die Frage zu geben, wie ein angemessener Umgang mit Deepfakes in der Schweiz aussehen kann. Dabei wird neben der näheren Untersuchung auf Deepfake-Technologien auf weitere fünf Bereiche fokussiert, die für eine vertiefte inhaltliche Auseinandersetzung mit Deepfakes essenziell sind: die Wahrnehmung von Deepfakes, der Umgang mit Deepfakes im Journalismus, Deepfakes im Recht, Deepfakes in der Politik sowie Deepfakes in der Wirtschaft. Auf diese Weise soll fundiertes Orientierungswissen für die Bevölkerung und für Akteure in Politik, Wirtschaft und anderen Gesellschaftsbereichen bereitgestellt werden, das idealerweise in besser reflektierte (politische) Entscheidungen mündet.

Trotz der gesellschaftlich aufgeladenen Debatte, in der Deepfakes vorwiegend kritisch betrachtet werden, möchte die vorliegende Studie zwischen Potenzialen und Herausforderungen vermitteln und einen möglichst holistischen Zugang zum Thema pflegen.

### 1.2.1. Fragestellung der Studie

Die Fragestellung der Studie ist in verschiedene thematische Fragekomplexe untergliedert.

*Ist- und Trendanalyse*

Forschungsfrage (FF) 1.1: Welche technischen Möglichkeiten existieren gegenwärtig und sind in den nächsten Jahren absehbar, um synthetische oder manipulierte Inhalte zu erzeugen?

FF 1.2: Welche (technischen) Möglichkeiten existieren, um sowohl echte als auch manipulierte Inhalte zu erkennen und zu kennzeichnen?

*Wahrnehmung von Deepfakes in der Schweizer Bevölkerung*

FF 2.1: Welche Erfahrungen haben Menschen in der Schweiz bislang mit Deepfakes?

FF 2.2: Werden Deepfakes in der Schweizer Bevölkerung eher als Chance oder als Risiko wahrgenommen und welche Faktoren beeinflussen diese Wahrnehmung?

FF 2.3: Sind Menschen in der Schweiz in der Lage, Deepfakes zu erkennen, und welche Faktoren beeinflussen diesen Prozess?

*Deepfakes im Recht*

FF 3.1: Können Deepfakes grundrechtlichen Schutz geniessen (etwa unter der Meinungs- oder Kunstfreiheit) und kann ein Urheberrecht an Deepfakes bestehen?

FF 3.2: Welche Rechtsvorschriften werden ggf. durch Deepfakes verletzt und welche rechtlichen Möglichkeiten dagegen können ergriffen werden?

FF 3.3: Welche Rolle könnten Deepfakes vor Gericht spielen und welche Lösungsansätze bieten sich zur Adressierung allfälliger Herausforderungen dabei an?

FF 3.4: Reicht die bestehende Gesetzgebung, um Individuen z.B. vor Mobbing und Erpressung und Unternehmen z.B. vor Betrug oder Rufschädigung zu schützen? Inwieweit besteht weiterer Regulierungsbedarf?

*Deepfakes im Journalismus*

FF 4.1: Welche Strategien werden im Journalismus zur Identifikation von Deepfakes angewendet und zu welchen konkreten Anpassungen von Arbeitsprozessen und Routinen führt dies in Redaktionen von Schweizer Medienorganisationen aktuell und in der Zukunft?

FF 4.2: Wie werden die Herausforderungen, welche Deepfakes für den Journalismus mit sich bringen, in der Ausbildung von Medienschaffenden aktuell und in Zukunft thematisiert?

FF 4.3: Wie sind Medienorganisationen auf Fälle vorbereitet, in denen Journalistinnen und Journalisten selbst von Deepfakes betroffen sind?

*Deepfakes in der Politik*

FF 5.1: Welche Rolle spielen Deepfakes gegenwärtig und künftig in der Schweizer Politik?

FF 5.2: Wie können die von Deepfakes in der Politik ausgehenden Herausforderungen adressiert werden?

FF 5.3: Welche Handlungsoptionen bieten sich der Politik und anderen Akteuren?

*Deepfakes in der Wirtschaft*

FF 6.1: Welche Herausforderungen durch Deepfakes sind im Hinblick auf Unternehmen zu erwarten?

FF 6.2: Welche Chancen bringen Deepfake-Technologien mit sich?

FF 6.3: Wie können sich Unternehmen vor Deepfakes schützen?

*Empfehlungen*

Auf Basis der vorangegangenen Kapitel wurden schliesslich Empfehlungen für (politische) Entscheidungstragende formuliert, wie ein verantwortungsbewusster Umgang mit den Chancen und Herausforderungen von Deepfakes aussehen kann.

## 1.3. Begriffsklärung

Die Debatte um Deepfakes und manipulierte Inhalte hat zur Verwendung einer Reihe von Begriffen geführt, die weiterer Spezifizierung bedürfen. Mit dem Aufkommen generativer KI-Technologien[1] wurde das Feld zuletzt noch unübersichtlicher. Auch, dass für die im Folgenden diskutierten Begriffe keine Legaldefinitionen existieren, erschwert die Einordnung. Daher werden nachfolgend zentrale Begriffe der Studie geklärt, um zu einer möglichst einheitlichen und klaren Begrifflichkeit zu gelangen.

**Deepfake:** Ein Deepfake ist ein mithilfe von KI-Techniken synthetisierter oder manipulierter Audio-, Bild- bzw. Videoinhalt, der authentisch wirkt, es aber nicht ist. Ein Deepfake zeigt häufig einen real existierenden Menschen, der etwas sagt oder tut, was er nie gesagt oder getan hat. Deepfakes können allerdings auch ohne menschlichen Bezug auskommen und beliebige Objekte, Situationen (bspw. synthetisierte oder manipulierte Naturkatastrophen) oder andere Lebewesen zum Gegenstand haben.

Wir begreifen Deepfakes als *dual-use*: Je nach Nutzung und Kontext können sie einen nützlichen oder einen schädigenden, einen erwünschten oder unerwünschten Effekt auf bestimmte Menschen, Gruppen oder Organisationen entfalten. Dieser Effekt kann intendiert oder auch nicht intendiert sein. Eine häufige Nutzungsform von Deepfakes ist ihr Einsatz zum Zwecke der Verbreitung von *Desinformation* und *Misinformation*; diese Begriffe werden unten näher definiert.

Über die Einordnung von Large Language Models (LLMs) als Deepfake-Technologie besteht derzeit noch Unklarheit. Angesichts der Möglichkeit, die Sprech- und Schreibweise eines Menschen mittels moderner Chatbots kopieren zu können, werden derartige Ergebnisse teilweise als Deepfake-Text bezeichnet. Weil sich dieser Sprachgebrauch allerdings noch nicht durchgesetzt hat, ordnen wir diese nicht in die Kategorie von Deepfakes ein. Stattdessen bezeichnen wir diese Texte als *synthetisierter Text* oder *KI-generierter Text*.

---

[1] Unser Verständnis von künstlicher Intelligenz folgt der Definition in Christen et al. (2020, 74): «Künstliche Intelligenz bezeichnet den Versuch, Verstehen und Lernen mittels eines Artefakts nachzubilden, wobei in erster Linie auf Denken bzw. Handeln fokuusiert sowie ein rationales Ideal bzw. eine Nachbildung menschlicher Fähigkeiten angestrebt wird.» Sowie: «KI-Technologie bezeichnet einzelne, in Computer implementierbare Funktionen für die Erreichung von künstlicher Intelligenz (z.B. maschinelles Lernen).»

**Cheapfake (oder Shallow Fake):** Um einen Cheapfake (oder Shallow Fake) handelt es sich bei Inhalten, die mittels Methoden erstellt wurden, die nicht auf KI-Technologien basieren (klassisches Photoshoppen, das Neuzusammenschneiden bestehenden Videomaterials oder das verlangsamte Abspielen eines Videos u.v.m.).

**Deepfake-Technologie(n):** Deepfakes können mit verschiedenen KI-Technologien hergestellt werden. Zu diesen Technologien zählen insb. Generative Adverserial Networks (GANs) und Generative KIs, wie z.B. Diffusionsmodelle. Insofern verwenden wir «Deepfake-Technologie» als Oberbegriff für alle KI-basierten Technologien, mittels derer ein Deepfake produziert werden kann. LLMs zählen wir aufgrund der o.g. Gründe nicht zu Deepfake-Technologien.

**Echtheit:** Mit Echtheit bezeichnen wir, ob eine Tatsache mit ihrer Darstellung, z.B. in Form eines Medieninhalts, übereinstimmt. Echtheit ist bei der Diskussion von Deepfakes einer von vielen Faktoren, die je nach Deepfake, Kontext, Zweck usw. anders zu gewichten sind. Wenn z.B. einer Person Worte oder Handlungen mittels Deepfake-Technologien unterstellt werden, welche diese Person nie gesagt oder getan hat, handelt es sich dabei um keinen echten Inhalt, sondern um eine Fälschung (s.u.), die mitunter rechtliche Konsequenzen nach sich ziehen kann (vgl. 4.2). Andererseits ziehen satirische Deepfakes ihren Unterhaltungswert gerade daraus, dass sie keinen echten Inhalt darstellen, sondern durch bewusste Überspitzung eine meinungsgefärbte Interpretation echter Sachverhalte bilden (vgl. 4.1.1). Ebenso basieren andere Formen unterhaltender Deepfakes, etwa das Schneiden des eigenen Gesichts in existierendes Filmmaterial, auf der Veränderung echter Inhalte.

**Original:** Als «Originalität» oder «original» definieren wir im technischen Sinne einen digitalen Inhalt, welcher erstmalig als solcher erstellt oder gespeichert wurde. Somit kann ein Original sowohl einen «echten» Inhalt zeigen als auch einen unechten Inhalt. Im Gegensatz zur «Echtheit» kann auch ein unechter Inhalt (z.B. eine Kopie oder eine Abzeichnung eines Kunstwerkes) ein «Original» sein. Es wird damit lediglich belegt, dass der erstmalige Inhalt nicht verändert wurde.

**Fälschung:** Eine Fälschung verstehen wir im Kontext von Deepfakes als einen Inhalt, der in Täuschungsabsicht einen echten oder originalen Inhalt nachbildet. Dabei ist es unerheblich, ob Originalmaterial manipuliert wurde oder ob ein Inhalt neu erstellt wurde. Entscheidend ist die Absicht, die Rezipienten zu täuschen, indem diese die Fälschung für das Original oder den echten Inhalt halten. Auch ein gefälschter Medieninhalt kann aus unterschiedlichen Gründen unrecht-

mässig (vgl. 4.2) oder vom Recht auf Meinungs-, Informations- und Kunstfreiheit abgedeckt sein (vgl. 4.1.1).

**Imitation:** Bei einer Imitation handelt es sich ebenfalls um einen Inhalt, der Eigenschaften eines Originals nachbildet. Anders als bei einer Fälschung liegt bei einer Imitation allerdings nicht zwingend eine Täuschungsabsicht vor.

**Manipulierte Medieninhalte:** Von manipulierten Medieninhalten sprechen wir, wenn ein originaler Inhalt mittels KI-Technologien so verändert wurde, dass es nicht mehr dem Original entspricht.

**Synthetische Medieninhalte:** Als synthetisch bezeichnen wir Medieninhalte, die mittels KI-Technologien, z.B. durch Eingabe von Textbefehlen in natürlicher Sprache (Prompts) anhand von zugrunde liegenden Trainingsmaterialien, erstellt wurden. Dabei kann es sich um Text-, Audio-, Bild- oder Videoinhalte handeln. Weil diese Kategorie auch Texte einschliesst, handelt es sich um eine Oberkategorie für jegliche, mit KI-Technologien erstellte Inhalte. Sofern auch manipulierte Inhalte diskutiert werden, ist im Bericht von «synthetischen» und «manipulierten Inhalten» die Rede.

**Irreführende Inhalte:** Als irreführenden Inhalt bezeichnen wir Text-, Audio-, Bild- und Videoinhalte, die dem Betrachter als authentisch erscheinen, es aber nicht sind. Insofern fungiert der Begriff synonym zu «Deepfake». Dabei ist es unerheblich, ob eine Irreführungsabsicht vorliegt oder nicht. Eine Irreführung kann also sowohl von einem bewusst zur Irreführung produzierten Inhalt ausgehen als auch von Inhalten, die z.B. zu Unterhaltungszwecken produziert wurden.

**Desinformation:** Desinformationen sind nachweislich falsche oder irreführende Informationen, die intentional aus politischen oder wirtschaftlichen Gründen verbreitet werden. Irrtümer bei der Berichterstattung, Satire und Parodien oder eindeutig gekennzeichnete parteiliche Nachrichten oder Kommentare sind keine Desinformation (Kalsnes u.a. 2021; HLEG 2018; European Commission 2018: 4; Tandoc Jr u.a. 2019).

**Fehlinformation (Misinformation):** Unter Fehlinformationen verstehen wir Informationen, «die faktisch nicht der Wahrheit entsprechen, jedoch vom Sender, der die Information verbreitet, für wahr gehalten werden» (Johann und Wagner 2018, 102). Insofern unterscheidet sich Fehlinformation von Desinformation hinsichtlich des Fehlens der Schädigungsabsicht.

**«Fake News»:** Der Begriff «Fake News» wurde zu Beginn der jüngsten Debatten rund um Desinformation häufig verwendet, um die vorsätzliche Täuschung der Öffentlichkeit mit nachweislich falschen Nachrichtenmeldungen durch Me-

dienorganisationen zu beschreiben (Kalsnes u.a. 2021), er wird inzwischen jedoch aufgrund seiner inhaltlichen Unschärfe und politischen Aufladung gemieden (European Commission 2018: 10; Wardle/Derakhshan 2018: 5). Auch wir vermeiden den Begriff nach Möglichkeit, greifen ihn aber stellenweise in Guillemets doch auf, wenn es der entsprechende Kontext gebietet (bspw. Diskussion der einschlägigen Forschungsliteratur).

## 1.4. Methodologie

Das Projektkonsortium vereinte interdisziplinäre Perspektiven, um die o.g. Forschungsfragen zu beantworten. Dabei wurde auf einen Methodenmix zurückgegriffen, der in Abschnitt 1.4.1 überblickshaft vorgestellt wird. Detaillierte Informationen zu der jeweiligen Methodik können den entsprechenden inhaltlichen Kapiteln entnommen werden. Die Fachexpertise des Projektteams wird in Abschnitt 1.4.2 vorgestellt.

### 1.4.1. Arbeitsschritte und Methoden

1. **Dokumenten- und Literaturanalyse:** Ein zentrales methodisches Instrument der Studie ist die Literaturanalyse. Dies schliesst die Aufarbeitung des Stands von Deepfake-Technologien und die bibliometrische Analyse in der Ist- und Trendanalyse in Kapitel 2 sowie die vertiefenden Analysen in den folgenden Kapiteln ein.
2. **Eigene technische Tests:** Zur Untersuchung der Leistungsfähigkeit von Deepfake-Detektoren führten wir im Rahmen der Ist- und Trendanalyse (2.7) eigene Tests mit heute verfügbaren Deepfake-Video-Detektoren durch.
3. **Analyse der Medienberichterstattung:** Neben der Analyse von Dokumenten und Literatur kam die Analyse der Medienberichterstattung in Kapitel 7 zum Einsatz, um Chancen von Deepfakes zu identifizieren.
4. **Expertinnen- und Experteninterviews:** Insbesondere im Rahmen von Kapitel 5 (sowie 2) wurden Experteninterviews durchgeführt.
5. **Bevölkerungsumfrage:** Zur Untersuchung der Wahrnehmung von Deepfakes in Kapitel 3 wurden zwei – im Hinblick auf Alter, Geschlecht, Sprachregion und Bildung repräsentative – Befragungen durchgeführt, in denen neben einer generellen Befragung jeweils ein Onlineexperiment eingebaut wurde.

Die Studie ist in acht inhaltliche Kapitel unterteilt. Das erste inhaltliche Kapitel 2 dient im Rahmen einer Ist- und Trendanalyse der Einführung in das Themengebiet. Nach einer Einführung in die Grundlagen von Deepfake-Technologien (2.1, 2.2, 2.3.1.5 und 2.5) werden technische Gegenmassnahmen in Form von Präventions- (2.6.1) und Erkennungsmassnahmen (2.6.2) diskutiert sowie die Ergebnisse unseres Detektor-Tests vorgestellt (2.7). Abschliessend folgt die bibliometrische Analyse des Publikationsaufkommens zum Thema Deepfakes (2.8). An die Ist- und Trendanalyse schliessen sich die vertiefenden Analysen in ausgewählten Themenfeldern an. Kapitel 3 widmet sich dem Thema der Wahrnehmung von Deepfakes in der Schweizer Bevölkerung. Nach der Einführung in Theorie und Forschungsstand (3.1) werden die methodische Vorgehensweise (3.2), Resultate (3.3) sowie Hauptbefunde (3.4) vorgestellt. Kapitel 4 zu Deepfakes im Recht diskutiert den Schutz *von* Deepfakes (4.1), den Schutz *vor* Deepfakes (4.2), die Frage nach Deepfakes vor Gericht (4.3) sowie öffentlich-rechtliche Vorgaben (4.4) und Regulierungsmöglichkeiten von Deepfakes (4.5). In Kapitel 5 werden Deepfakes im Journalismus untersucht. Dazu wird zunächst in Theorie- und Forschungsstand eingeführt (5.1), die methodische Vorgehensweise vorgestellt (5.2) und schliesslich werden die Resultate (5.3) und Hauptbefunde (5.4) diskutiert. In Kapitel 6 werden Deepfakes in der Politik untersucht. Nach einer Einführung in den Forschungsstand (6.1) werden zunächst die Ergebnisse unserer Umfrage im Schweizer Parlament und der Bundesverwaltung vorgestellt (6.2) und anschliessend Szenarien zu Deepfakes in der Politik (6.3) erörtert. In Kapitel 7 werden Deepfakes in der Wirtschaft untersucht. Zunächst werden Herausforderungen (7.1) diskutiert und im Anschluss Chancen (7.2). Anschliessend werden Szenarien zu Deepfakes in der Wirtschaft erörtert (7.3) und Massnahmen zum Schutz und zur Schadensbegrenzung (7.4) aufgezeigt. Das finale Kapitel 8 enthält die Empfehlungen und Kapitel 9 die Schlussfolgerungen der Studie (9).

### 1.4.2. Projektkonsortium

Die Projektmitarbeiterinnen und -mitarbeiter des Konsortiums waren entsprechend ihrer jeweiligen Expertise für unterschiedliche Teile der Studie federführend zuständig.

Das **Fraunhofer-Institut für System- und Innovationsforschung ISI** in Karlsruhe, Deutschland, ist eine von mehr als 80 Forschungseinrichtungen der Fraunhofer-Gesellschaft. Es hat einen Fokus auf sozioökonomische Forschung, Vorausschau, Evaluation, Wirkungsanalysen und Technologiefolgenabschätzung

sowie Politikberatung. Das ISI ist Mitglied in der European Technology Assessment Group (ETAG) und im globalTA Network. Für die Durchführung der Studie waren Wissenschaftler aus dem Geschäftsfeld «Informations- und Kommunikationssysteme» im Competence Center «Neue Technologien» zuständig. Die Mitarbeitenden des ISI-Teams bringen Expertise aus verschiedenen Disziplinen mit: Technik, Wirtschaftsinformatik sowie Sozial- und Politikwissenschaften.

Das Fraunhofer ISI war neben der Projektkoordination und der Zusammenführung der Inputs aller Projektpartner für die Erarbeitung der Einleitung, der Ist- und Trendanalyse, der Kapitel zu Deepfakes in der Politik und Deepfakes in der Wirtschaft sowie der Schlussfolgerungen zuständig. Bei der Ausarbeitung der Empfehlungen unterstützte es das Projektteam des Instituts für Europarecht.

Das **Institut für Europarecht der Universität Freiburg i. Ue.** gehört zu den führenden Forschungseinrichtungen für Datenschutz in der Schweiz. Seit über zehn Jahren richtet das zweisprachige (d./fr.) Institut, zusammen mit dem Eidgenössischen Datenschutzbeauftragten (EDÖB), den Schweizerischen Datenschutzrechtstag sowie Weiterbildungen für Berufspraktikerinnen und -praktiker im Datenschutzrecht aus.

Das Projektteam des Instituts für Europarecht war federführend für die juristische Bewertung der Anwendungsfelder und die Erarbeitung der Empfehlungen (mit Unterstützung des Fraunhofer ISI) zuständig. Ausserdem wirkte das Projektteam des Instituts für Europarecht unterstützend bei den anderen Arbeitsschritten, für die rechtliche Expertise notwendig war.

Das **Departement für Kommunikationswissenschaft und Medienforschung (DCM) der Universität Freiburg i. Ue.** beschäftigt sich mit der Erforschung und Reflexion der vielfältigen Phänomene, Prozesse und Strukturen medienvermittelter öffentlicher Kommunikation in der Schweiz sowie im internationalen Vergleich. Der Lehrstuhl für «Medienstrukturen und Governance» beschäftigt sich mit dem Schweizer Mediensystem und Mediensystemen im internationalen Vergleich, Medienpolitik, Medienregulierung und Media Governance, Medienorganisationen und Journalismus sowie Medienökonomie und Kritischer Politischer Ökonomie. Dabei interessiert insbesondere, wie Digitalisierung, Kommerzialisierung und Globalisierung die Rolle der Medien in der demokratischen Gesellschaft verändern und welche medienpolitischen Optionen für die Aufrechterhaltung eines leistungsfähigen Mediensystems bestehen. Das Projektteam des DCM war für die Teilstudie zu Deepfakes im Journalismus zuständig (Kapitel 5).

Das **fög – Forschungszentrum Öffentlichkeit und Gesellschaft** ist ein auf Medienforschung spezialisiertes Zentrum der Universität Zürich. Seit seiner Gründung im Jahr 1997 legt das Zentrum den Schwerpunkt seiner Forschung auf den digitalen Strukturwandel der Öffentlichkeit, insbesondere auf den Wandel der Medien und dessen Konsequenzen für die Gesellschaft. Mit dem *Jahrbuch Qualität der Medien* untersucht das fög, das ein interdisziplinäres Team von 15 Mitarbeitenden beschäftigt, seit über zehn Jahren die Entwicklung des Mediensystems und des Nutzungsverhaltens systematisch und auf empirischer Basis. Das Forschungszentrum ist national und international stark vernetzt und unterhält langjährige Forschungskooperationen mit namhaften Institutionen, darunter der University of Oxford. Das Projektteam war für die Teilstudie zur Wahrnehmung von Deepfakes in der Schweizer Bevölkerung zuständig (Kapitel 3).

# 2. Ist- und Trendanalyse

*Frank Ebbers, Murat Karaboga, Greta Runge & Michael Friedewald*

In diesem Kapitel wird der technische Hintergrund zur Erstellung von Deepfakes beschrieben. Nach einer kurzen historischen Einführung in die Geschichte von Deepfake-Bildern und -Audios (2.1) wird auf Technologien zur Erstellung von Bild- und Videomaterial (2.2 und 2.3) sowie von Audio (2.3.1.5) und Text (2.5) eingegangen. Danach folgt eine Betrachtung von Technologien zur Prävention und Erkennung von Deepfakes (2.6) und die Vorstellung eigener Testergebnisse zu Deepfake-Detektoren (2.7). Abschliessend werden die Ergebnisse der bibliometrischen Auswertung wissenschaftlicher Publikationen zu Deepfakes vorgestellt (2.8).

## 2.1. Historischer Rückblick zu den Grundlagen von Deepfakes

### 2.1.1. Digitale Manipulation von Bildern und Videos

Als Computer in den 1970er-Jahren mit grafischen Benutzeroberflächen ausgestattet wurden, wurden auch die ersten Anwendungen für die Bildbearbeitung entwickelt. Als bekanntestes Beispiel gilt Photoshop, welches in den 1990er-Jahren populär wurde. So erhielt erstmals eine breite Masse an Menschen die Möglichkeit, Bilder auf vergleichsweise einfache Weise digital zu manipulieren. Die begrenzten Möglichkeiten der ersten Versionen der Software wurden seit seiner Veröffentlichung sukzessive erweitert, und damit auch die Ausgereiftheit der Bearbeitungsergebnisse (Farid/Schindler 2020: 9).[2]

Die Manipulation von Videomaterial wurde aufgrund steigender Rechenkapazitäten ebenfalls im Laufe der 1990er-Jahre möglich. Ebenso wurde das Synthetisieren von Gesichtsinformationen möglich. Als prominentes Beispiel zählt ein Video aus dem Jahr 1997, in dem die Gesichtszüge des verstorbenen US-Präsidenten John F. Kennedy manipuliert wurden (Bregler et al. 1997). Mithilfe der hierzu eingesetzten Software namens «Video Rewrite» konnten die Mund-

[2] Seit 2023 unterstützt Photoshop in der Standardversion generative Füllung bzw. KI-gestützte Bilderzeugung.

bewegungen in einem vorhandenen Video so verändert werden, dass sie etwas anderes sprechen. Es war das erste Programm, welches zeitgleich Gesichter interpretieren, Audio aus Text synthetisieren und Lippen im 3D-Raum modellieren konnte. Hochwertige Videobearbeitung blieb vorerst Akademikern auf dem Gebiet der experimentellen Bildverarbeitung und den Filmstudios mittels CGI vorbehalten. Mit weiter steigender Rechenleistung wurden schliesslich auch die Ergebnisse der Filmstudios immer besser. Ein bekanntes Beispiel ist der Gewinner des Academy Award 2009 für die besten visuellen Effekte: *Der seltsame Fall des Benjamin Button*. Während des gesamten Films werden computergestützte Manipulationen am Gesicht des Schauspielers Brad Pitt vorgenommen, um die Illusion einer umgekehrten Alterung zu erzeugen (Song 2019).

Den Boom von Deepfake-Anwendungen löste allerdings erst der Rückgriff auf KI-Technologien in Gestalt neuronaler Netze aus. Während erste GAN-basierte Anwendungen im Jahr 2014 auftauchten und noch recht gut als Fälschungen erkennbar waren, gelang dem Grafikkartenhersteller Nvidia 2017 mittels Rückgriffs auf das stufenweise Training von neuronalen Netzen (mit einer grossen Menge an Trainingsmaterial) ein grosser technologischer Qualitätssprung, wodurch die Generierung von Videos in höherer Qualität möglich wurde (Schreiner 2022). Denn zunächst war die Auflösung auf ein paar Hundert Pixel begrenzt, weil weniger Pixel zu weniger Fehlern führen und das Endprodukt schwerer als Fälschung erkannt werden kann. Kaum ein Jahr später wurde es möglich, den Algorithmen Vorgaben (sog. «Style-Transfer»), wie zum Beispiel «dunkle Haare» oder «Lächeln», zu machen (ebd.). 2015 kam mit der Face-Swapping-Technologie die Möglichkeit auf, Gesichter in Echtzeit zu ersetzen (Breithut 2015). Diese Technologie wurde über die Jahre stetig verbessert, sodass es im Jahr 2019 möglich wurde, die Gesichter in Livevideos in Echtzeit auszutauschen, ohne vorher Trainingsmaterial zu dem Gesicht analysiert zu haben (Schreiner 2022). Die technologischen Fortschritte führten über die Jahre zur Veröffentlichung einer Vielzahl von Anwendungen zur Produktion von Deepfakes (z.B. das DeepFaceLab) (Mirsky und Lee 2022a, S. 1). Jüngste Fortschritte ermöglichen die Erschaffung von Deepfakes mit Auflösungen von bis zu 2048 x 1024 Pixeln. Zudem können generative KI-Anwendungen, wie Dall-E 2 (OpenAI 2022), Bilder rein aus einer Beschreibung in natürlicher Sprache (sog. Text-Prompts) erstellen (Schreiner 2022; Mirsky/Lee 2022: 10). Je besser, also insb. detaillierter, ein Text-Prompt bzw. eine Kette von aufeinanderfolgenden Text-Prompts ist, umso authentischer kann ein Deepfake-Inhalt werden. Doch kann die Formulierung derartiger, ausgefeilter Text-Prompts schwierig sein. Einen Weg zur Formulierung zielführender Text-Prompts bietet wiederum ChatGPT. Dort können Nutzende einen gewünschten Inhalt in ihrer Sprache beschreiben und den Chatbot

dazu auffordern, dies in Form eines Text-Prompts auszugeben, sodass die Qualität des Resultats gesteigert wird (Foley 2023; Gewirtz 2023). Abbildung 2 zeigt ein durch Dall-E 2 generiertes Bild, welches als Prompt die Textbeschreibung «ein Astronaut, reitet ein Pferd, in fotorealistischem Stil» erhalten hat.

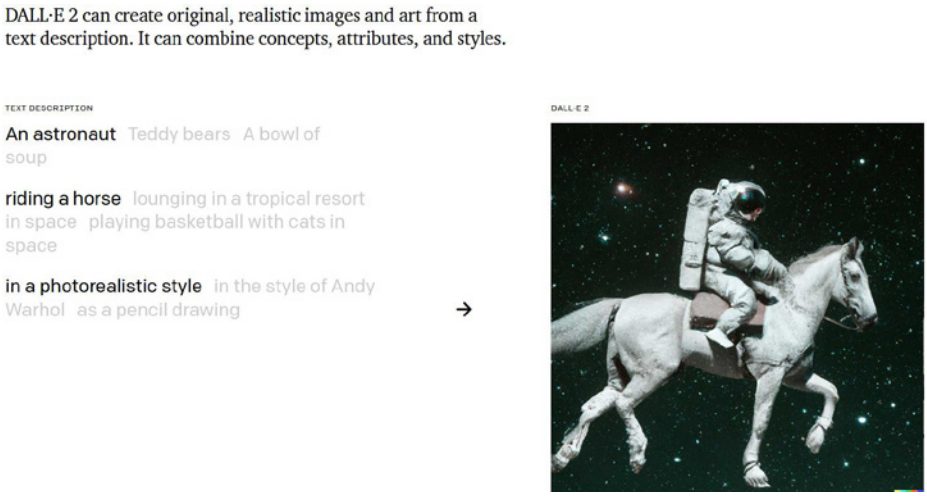


**Abbildung 2:** Beispiel eines aus Text generierten Bildes von DALL·E 2 (OpenAI 2022)

### 2.1.2. Klonen und Synthetisieren von Stimmen

Erste Software-basierte Versuche zum Klonen bzw. Synthetisieren von Stimmen begannen vor Jahrzehnten, allerdings galt die Stimme lange Zeit noch als sehr schwer zu fälschen. Zum einen, da Menschen eine sehr ausgeprägte Sensibilität für die Stimme haben und selbst kleinere Artefakte darin zu erkennen vermögen. Zum anderen, weil die Software zur Manipulation von Stimmsignalen nicht ausreichte, um genau diese Sensibilität zu täuschen (Napolitano 2020: 4).

Die Sprachsynthetisierung wurde als Text-to-Speech (TTS) bezeichnet. TTS-Algorithmen sind in der Lage, Text in gesprochene Wörter umzuwandeln und ermöglichen es Computern, mittels Sprache mit den Nutzenden zu interagieren. Während diese anfangs noch künstlich bzw. «roboterhaft» klangen, haben sich auch hier durch das Vorhandensein einer Vielzahl von Audiomaterial zum Training die Ergebnisse stark verbessert (van Huijstee u.a. 2021: 12). Zudem werden heutzutage auch im Bereich der Stimmsynthetisierung GANs eingesetzt, um möglichst authentisch klingende Stimmen zu erstellen. Im Bereich der

Stimmgenerierung ist die Entwicklung heute somit bereits fortgeschrittener als für Videomaterial und der Unterschied zwischen einer echten und einer synthetischen Stimme ist für den Durchschnittsmenschen kaum mehr wahrnehmbar (Martin 2020). Zudem sind auch Computer zunehmend weniger in der Lage, jede synthetisierte Stimme als solche zu erkennen (Wenger u.a. 2021: 235). Neben der Erschaffung von Fakes ist TTS heute zu einem Standard der alltäglichen Unterhaltungselektronik geworden und findet sich vor allem in Sprachassistenten und Navigationssystemen.

## 2.2. Technologien zur Fälschung und Synthetisierung von Bild- und Videomaterial

Foto- und Videomaterial werden mit ähnlichen Technologien gefälscht. Ein Video besteht technisch aus vielen Bildern, welche hintereinander gesetzt werden. Zur Erstellung eines Deepfakes-Videos wird jedes Bild einzeln manipuliert und entsprechend aneinandergereiht.

Drei Faktoren haben massgeblich für die rasante Entwicklung von Deepfakes geführt und sind die grundlegende Voraussetzung für den Einsatz von *Deep-Learning* (automatisiertes Selbstlernen von Computern). Diese sind (van Huijstee u.a. 2021: 7):

1. Wissenschaftler haben Algorithmen entwickelt, die automatisch Gesichtsmerkmale, wie die Position von Augenbrauen und Nase, in Bildern abbilden können.
2. Grosse Mengen an audiovisuellem Material sind dank Video- und Foto-Sharing-Plattformen (bspw. YouTube) verfügbar.
3. Die Entwicklung von Algorithmen sowie die Zunahme der Kapazitäten für Bildforensik, die eine automatische Erkennung von Fälschungen ermöglicht.

In der Praxis kommen meist mehrere verschiedene neuronale Netze zum Einsatz (Mirsky/Lee 2022: 7), wobei zwei spezifische KI-Ansätze sehr prominent sind: Generative Adversarial Networks (GANs) (siehe 2.2.1.) und Autoencoder (siehe 2.2.2.) (van Huijstee u.a. 2021: 7; Mirsky/Lee 2022: 6f.). Durch diese Methoden lassen sich auch ganze dreidimensionale Modelle einer Person erstellen und somit auch die Bewegung imitieren (sog. «3D morphable models») (Mirsky/Lee 2022: 7). Zum Zeitpunkt dieser Studie werden solche Modelle jedoch hauptsächlich für die Generierung von 3D-Avataren in Spielen oder bei Meta Platforms für das aufkommende Metaverse genutzt (van Huijstee u.a. 2021: 8).

### 2.2.1. Generative Adversarial Networks

GANs sind Computerprogramme, welche ein ähnliches, aber neuartiges Bild im Vergleich zu einem Trainingsset erzeugen können, indem sie eine Lernstrategie mit Rückkopplungsschleifen («Feedback Loop») verwenden (Mirsky/Lee 2022: 6f.). Die Programme bestehen aus zwei konkurrierenden Programmteilen: einem sog. «generativen Netzwerk» (g), das Inhalte durch die Analyse eines grossen Trainingsdatensatzes erzeugt. Im Fall von Deepfakes erkennt dieses generative Netzwerk anhand von Gesichtserkennungsalgorithmen gemeinsame Muster in Bildern und erschafft daraus neue ähnliche Inhalte. In Konkurrenz zum *generativen Netzwerk* steht das sog. «diskriminierende Netzwerk» (d), das Fälschungen auf der Grundlage forensischer Algorithmen erkennt. Hierzu prüft es, ob die neugenerierten Inhalte überzeugend «echt» sind oder dem Trainingsdatensatz ähneln. Jedes Mal, wenn das diskriminierende Netz eine Fälschung erkennt, nimmt das generative Netz dies zur Kenntnis und versucht, sein Ergebnis zu verbessern. GANs können zur Erstellung jeglicher Art von Inhalten, also auch von Tonaufnahmen und Text, eingesetzt werden (van Huijstee u.a. 2021: 79; Budhkar u.a.). Mathematisch ausgedrückt versuchen beide Teile eine Min-Max-Gleichung möglichst zu ihrem Vorteil zu füllen (Arjovsky/Bottou 2017: 2). Vereinfacht ergibt sich folgende Formel:

$$\min(d)\ \max(g) \rightarrow [\overline{\mu_d}(echt) - \overline{\mu_g}(1 - echt)]$$

Der Generator (g) möchte die Kosten der Gleichung maximieren max(g), während der Diskriminator diesen minimieren möchte min(d). stellt dabei den mittleren Erwartungswert dafür dar, dass ein Ergebnis echt ist. Diesen Wert versucht der Diskriminator möglichst gering zu halten, während der Generator versucht, den Wert möglichst hoch zu halten. Im Laufe vieler Iterationen spielen sich diese Min-Max-Funktionen gegenseitig aus.

Herausforderungen und Nachteile von GANs

Obwohl GANs grosse Bedeutung in der Forschung erfahren, haben sie noch einige recht abstrakte Nachteile und Herausforderungen, die zum Zeitpunkt dieses Berichts nicht vollständig gelöst sind (Google Developers 2022). Saxena/Cao (2022: 7 ff.) beschreiben hierzu drei übergeordnete Probleme:

**1. Kollaps (mode collapse)**

Der «Mode Collapse» beschreibt das Problem, dass sich der Generator auf die Einstellung seiner Gewichte fixiert. Dies hat zur Folge, dass der Generator nur noch sehr ähnliche Ergebnisse produziert (Saxena/Cao 2022: 7f.). Dieses Prob-

lem kann teilweise durch die sog. «Minibatch discrimination» adressiert werden. Dazu wird beim Training die Distanz von Datenpunkten zueinander gemessen und schliesslich aufsummiert und dem Diskriminator als zusätzliches Feedback gegeben (Salimans u.a. 2016: 1).

**2. Nicht-Konvergenz und Instabilität (non-convergence and instability)**

Problem Nummer zwei beschreibt die Fehleranfälligkeit aufgrund instabiler Trainingsmethoden. Denn durch das «unüberwachte Lernen» («unsupervised learning») ist ein verlässliches Training noch schwer.

Zudem ist es auch möglich, dass Generator und Diskriminator oszillieren. Wenn entweder der *d* oder *g* stärker wird, kann es zu Situationen kommen, in denen das System nicht mehr lernt. Dies führt zum sog. «Vanishing Gradients Problem» (Saxena/Cao 2022: 7). Dies sei eine der Hauptursachen dafür gewesen, dass tiefe neuronale Netze anfänglich nicht trainiert werden konnten (GI 2022). Hintergrund ist, dass je *tiefer* ein neuronales Netz wird, desto mehr Faktoren multipliziert werden müssen. Wenn diese während des Trainings kleiner als null sind, «dann multiplizieren sich mehrere Faktoren zu einer Zahl, die schnell gegen Null strebt, weswegen die Gewichtsänderungen in tiefen Schichten deutlich langsamer sind als die in höheren» (ebd.). Inzwischen gibt es allerdings mehrere Ansätze zur Beseitigung dieses Problems. Als erfolgreich stellte sich die Nutzung von mehr Rechenleistung heraus, was inzwischen effizient möglich ist. Zudem können nicht lineare Funktionen in neuen Netzarchitekturen genutzt werden (sog. ReLU-Aktivierungsfunktionen). Deren Ableitung ist für $x > 0$ exakt eins, sodass der Gradient nicht verschwinden kann («vanish») (GI 2022).

Eine Herausforderung bleibt auch dann noch, wenn ein Training abgeschlossen ist und ein Modell steht. Denn ab diesem Moment ist es nicht mehr möglich, Änderungen am Modell vorzunehmen (sog. «Inverting»). Neue Modelle erfordern daher immer wieder die Neuberechnung, was sowohl Zeit als auch Geld kostet (Gupta 2020). Die Kosten für das Training hängen hierbei signifikant von der Anzahl der Parameter ab. Das Training eines Sprachmodells *(Natural Language Processing)* mit 110 Millionen Parametern kostet zwischen 2500 und 50 000 US-Dollar pro Durchlauf. Bei 1,5 Milliarden Parametern sind es bis zu 1 600 000 US-Dollar, was bei der Notwendigkeit von mehreren Durchläufen zu Kosten in achtstelliger Höhe führen kann (Sharir u.a. 2020: 2).

**3. Fehlende Bewertungsmetriken (evaluation metrics)**

Problem Nummer 3 basiert auf der Frage, wie die Fähigkeit von GANs zur Erzeugung realistischer Daten am besten gemessen werden kann (Barua u.a.

2019: 1). Die meisten GANs beruhen auf einer Berechnung der nächstgelegenen Nachbarabstände (n-nearest neighbors) zwischen Stichproben aus den Originaldaten. Aus diesem Grund ist eine Erstellung von etwas «Neuem», das möglichst nicht in den Ausgangsdaten zu finden ist, schwer zu bewerten (Gupta 2020). Daraus resultiert auch das Problem der «Dichteschätzung». Man ist noch nicht in der Lage, die Genauigkeit der Dichte des erstellten Modells vorherzusagen. Somit müssen Metriken zur Dichte eines Modells («Thresholds») manuell festgelegt werden, um dann festzulegen, ob das erstellte Bild gut genug ist, um mit diesem im Generator weiterzuarbeiten (ebd.).

Fassbare Beispiele:

Konkrete und weniger abstrakte Schwierigkeiten haben GANS beim Einfärben von Graustufenbildern, da es schwierig ist, die Kolorierung von Farbbildern als eindeutig richtig oder falsch zu bewerten (Stieler 2021). GANs «neigen dazu, farblich möglichst ‹gemittelte Bilder› zu erzeugen – grosse Flächen also beispielsweise mit Brauntönen zu füllen, deren Farbwerte in der Mitte des Spektrums liegen und damit automatisch kleine Abweichungen erzeugen» (ebd.). Zudem haben sie Probleme, die realistischen Bewegungen von Haaren (z.B. bei Kopfbewegungen) darzustellen. Allerdings wird auch an der Überwindung dieser Herausforderungen geforscht (Metaphysic.ai 2022).

### 2.2.2. Autoencoder

Um das Gesicht einer Person in einem (Live-)Video zu ändern und die Gesichtsausdrücke und Mimik in Echtzeit anzupassen, werden Autoencoder eingesetzt. Dies geschieht in drei Schritten (van Huijstee u.a. 2021: 79):

1. Erkennung und Abgleich der Pose und des Gesichtsausdrucks in jedem einzelnen Bild des Ziel- und des Ausgangsvideos.
2. Lernen der Veränderung von Gesichtszügen einer bestimmten Person bei bestimmten Ausdrücken.
3. In Beziehung setzen dieser Veränderungen zueinander (z.B. gleichzeitige Veränderungen des Mundes und der Augenbrauen einer Person beim Lächeln).

Nach diesen Trainingsschritten ist der Autoencoder in der Lage, den Gesichtsausdruck in jedem beliebigen Bild der Zielperson anzupassen, indem es den Ausdruck in einem Ausgangsbild erkennt und übernimmt.

## 2.3. Sechs Techniken zur Erstellung von bildbasierten Deepfakes

Die eingesetzten Technologien können sowohl für komplette Fakes genutzt werden oder auch nur, um Teilbereiche eines Videos oder Bildes zu verändern («discrete deepfakes») (van Huijstee u.a. 2021: 8f.). Im Folgenden werden fünf gängige Techniken sowie eine neue, sechste Technik kurz vorgestellt. Verwandte Methoden zur Erschaffung von Cheapfakes werden an dieser Stelle nicht betrachtet, da es sich dabei um keine Deepfakes handelt.

### 2.3.1. Manipulation des Gesichtsausdrucks («facial reenactment»)

Hierbei werden bestimmte Teile eines Gesichts verändert, wobei die Identität der Zielperson erhalten bleibt. In Abbildung 3 ist zu erkennen, wie die Mimik eines Schauspielers auf die Zielperson (George W. Bush) übertragen wird («facial reenactment»). Diese Technik wird auch für die «visuelle Synchronisation» verwendet, bei der nur die Bewegung der Lippen einer Zielperson auf der Grundlage der Änderung des Tons oder unter Verwendung von Texteingaben angepasst wird (Mirsky/Lee 2022: 17). Mit diesen Techniken können beliebige Bereiche einer Person, z.B. deren Kopfhaltung, verändert werden.

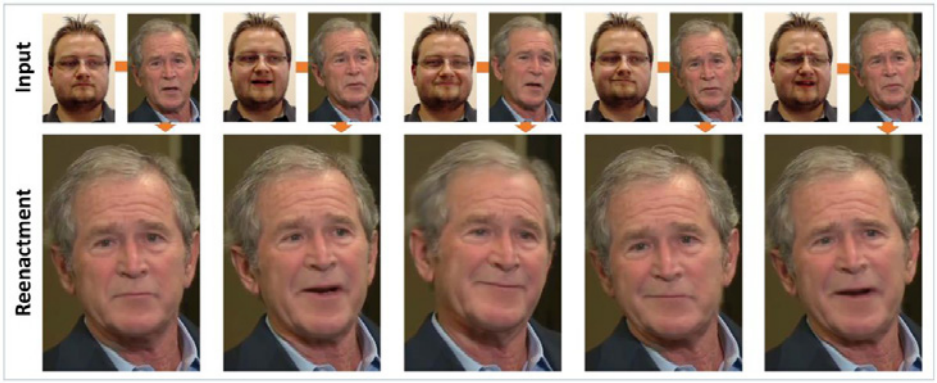


**Abbildung 3:** Facial-Expression-Manipulation > Georg W. Bush

### 2.3.2. Gesichts-Morphing («face-morphing»)

Das Ziel dieser Technik ist es, ein Bild oder Video zu erstellen, das aus zwei Gesichtern verschmilzt (Abbildung 4). Solche Techniken kommen beispielsweise in kriminellen Handlungen zum Einsatz, um authentische Ausweisdokumente (z.B. Pässe) zu erhalten, die von mehreren Personen zugleich verwendet werden können (Damer u.a. 2018: 1).


**Abbildung 4:** Photomorphing (Scherhag u.a. 2019: 302)

### 2.3.3. Gesichtsaustausch («face-swapping»)

Hierbei wird das Gesicht einer Zielperson durch das Gesicht des Ausgangsbildes oder -videos ersetzt (Abbildung 5).


**Abbildung 5:** Face-Swapping (Li u.a. 2019: 8)

### 2.3.4. Gesichtsgeneratoren

Mittels dieser Generatoren werden teils oder komplett neue Bilder von Personen erstellt, die so in der Welt nicht existieren. Hierzu werden die in 2.2.1. erörterten GANs verwendet. Sie werden zum Beispiel zur Animation von Charakteren in Videospielen oder Filmen oder in VR-Anwendungen als virtuelle Gesprächspartner genutzt (Thies u.a. 2016: 2387).

**Abbildung 6:** Mittels GAN generierte Fotos von Personen, die in der Realität nicht existieren (Schreiner 2022)

### 2.3.5. Ganzkörperpuppenspiel («full body puppertry»)

Mit Techniken zu Ganzkörperpuppenspielen kann die Pose eines Körperteils oder des gesamten Körpers einer Zielperson in einem Bild oder Video verändert werden. Als Vorlage kann ein vorhandenes Video dienen. Die full body puppertry wird jedoch als die komplizierteste Form der Deepfakes gesehen (Laufenburg 2021). Deshalb gibt es auch nur wenige echte Beispiele. In dem Modell «Everybody Dance Now» aus 2018 wird diese Technik angewendet (Abbildung 7). Der Algorithmus erkennt die Pose einer professionellen Tänzerin in einem Ausgangsvideo (oben links) und generiert ein Video (rechts), in dem die Zielperson die gleichen Tanzschritte aufführt. Informationen über den genauen Algorithmus fehlen. Jedoch monieren Dagar/Vishwakarma (2022: 223), dass dieser Algorithmus keine realistischen Körperhaltungen zeigt, insbesondere an den Gelenken. Ein weiteres Beispiel wurde von Gafni u.a. (2021) publiziert. Sie stellen eine Methode vor, die ein einzelnes Bild durch beliebige Videosequenzen reanimieren kann. Eine Bewertung durch Dagar/Vishwakarma (2022: 231) zeigt jedoch,

dass es keine kohärenten Ergebnisse liefert und die Kanten verpixelt sind. Im Jahr 2023 stellte ein anderes Forscherteam einen Algorithmus vor, welcher die Körperhaltung einer Person nach bereits neun Minuten Training ändern kann (Cheres und Groza 2023, S. 31850).


**Abbildung 7:** «Everybody can Dance»: Beispiel für die Technik «full body puppertry» (Chan u.a. 2019)

### 2.3.6. Aus Texteingaben generierte Deepfake-Videos («Text-to-Video»)

Ähnlich wie Bilder durch Texteingaben generiert werden können, ist dies auch für die Erstellung von Videos möglich («AI Video Generator» oder «Text-to-Video»-Generator). Letztendlich sind Videos nur eine Aneinanderreihung von Bildern. Die Herausforderung für den Algorithmus ist dabei, dass die Bilder aufeinander aufbauen müssen, um einen Zusammenhang darzustellen.

Prominente Beispiele für solche Anwendung sind zum Beispiel Metas «Make-A-Video» (Meta AI 2023) oder Google Imagen (Google Developers 2023), welche jedoch noch nicht öffentlich verfügbar sind. Mit dem Gen-2-Modell hat Runway ML (2023) ein vielbeachtetes Tool geschaffen, welches kostenlos genutzt werden kann – wenn auch mit einigen Einschränkungen. Es bietet laut den Entwicklern jedoch sehr viele feine Einstellungsmöglichkeiten und kann den gewünschten Videostil sehr gut wiedergeben (Esser u.a. 2023). Weiter gibt es Deepbrain AI (2023), welches sich auf virtuelle Avatare und Stimmen spezialisiert hat. Kaiber kann zusätzlich zu Text-to-Video auch Animationen mit unterschiedlichen Vorgaben aus Bildern oder Videos erstellen (Kaiber 2023). Allerdings sind die Preise für benötigte Credits relativ hoch (Wöbbeking 2023). Videos sind auf maximal

acht Minuten beschränkt. Animationen können jedoch in 4K aufgelöst werden (Kaiber 2023). Mit Pictory.ai (2023) lassen sich sogar Videos mit passenden Stimmen aus vorgefertigten Regieplänen und Blogeinträgen erstellen. Zusätzlich gibt es noch eine ganze Reihe weitere Programme, welche die Erstellung von Videos unterstützen (siehe z.B. Junghärtchen 2023).

Zum Zeitpunkt dieses Berichts erstellen diese Systeme grundsätzlich Deepfake-Videos noch in geringer Auflösung und mit geringer Länge. Es ist jedoch absolut wahrscheinlich, dass in naher Zukunft auch Full-HD und 4K-Videos möglich sein werden. So arbeitet der Grafikkartenhersteller Nvidia bereits an höheren Auflösungen (Koch 2023). Google Imagen «umgeht» dieses Problem damit, dass zuerst Videos mit einer Auflösung von 24 x 48 Pixel bei 3 Bildern pro Sekunde erstellt und die Bilder nachträglich mit einer anderen KI hochskaliert und Zwischenbilder generiert werden, sodass am Ende eine Auflösung von 1280 x 768 Pixeln, bei 24 Bildern entsteht (Ho u.a. 2022: 8). Blattmann u.a. (2023: 1) erstellten mittels eines Latent Diffusion Models (LDMs) einen Text-to-Video-Generator, welcher Videos mit 1280 x 2048 Pixel generieren kann. Der Fokus liegt hier aktuell auf «kreative Videoinhalte» und «simulierte Autofahrten» (Koch 2023). Ein LDM ist ein Deeplearning-Algorithmus zur Erstellung von hochauflösenden Videos. Ähnlich wie andere Diffusionsmodelle, beginnen die LDM auch mit zufälligem Rauschen und wandeln dieses schrittweise in realistische Bilder. Jedoch basiert dies bei LDMs nicht auf den Pixeln eines Bildes, sondern auf einer kodierten Darstellung des Bildes (Neto 2023). Dies benötigt weniger Rechenleistung. Das LDM wird zuerst mit Bildern trainiert und nachträglich wird eine Zeitebene in das Training eingefügt (Blattmann u.a. 2023: 1).

Über das tatsächliche Modell oder die zugrunde liegenden Daten ist jedoch wenig bekannt. Über «Make-A-Video» kann man auf der offiziellen Webseite lesen, dass das System «annotierte Bilder nutzt, um zu lernen, wie die Welt aussieht» und «nicht annotierte Videos, um zu verstehen, wie sich die Welt bewegt» (übersetzt aus Meta AI 2023). Auch in einem Forschungspapier zu «Make-A-Video» sind keine Informationen über die benutzten Trainingsbilder und Videos zu finden (Singer u.a. 2022: 6). Google Imagen wurde mit «problematischen Bildern» trainiert (Ho u.a. 2022: 15). Jedoch versucht das Unternehmen aktuell, diese aus dem Trainingssatz zu entfernen (Nordenbrock 2022).

Abschliessend lässt sich festhalten, dass es zum aktuellen Zeitpunkt kein KI-Gesamtpaket gibt, welches ein komplettes Video mit Sprecheranimation und Stimmsynthetisierung erstellen kann. Somit müssen mehrere kostenpflichtige Dienste verbunden werden, wodurch es neben hohen finanziellen Kosten (dreistellige monatliche Nutzungsgebühren) auch zu Qualitätsverlusten während der Zusammenstellung kommt (Junghärtchen 2023).

## 2.4. Technologien zum Klonen der Stimme

Die Begrifflichkeiten im Englischen zur Beschreibung von Deepfake-Audios sind vielfältig und reichen von «voice cloning», über «voice swapping» zu «speech synthesis» (van Huijstee u.a. 2021: 12), wobei die ersten beiden Begriffe nur im Kontext von Deepfakes zur Sprache kommen (Napolitano 2020: 4).

Der technische Fortschritt beim Klonen von Stimmen hat in den letzten Jahren stark zugenommen. Während Stimmen früher noch sehr unnatürlich und roboterhaft klangen, zeigen Amezaga/Hajek (2022: 23), dass bereits ein handelsüblicher Laptop und wenige Sekunden Referenz-Audiomaterial ausreichen, um eine Stimme zu klonen. Zudem ist es heutzutage einfach, dieses Referenz-Audiomaterial zu bekommen – beispielsweise von Vorträgen einer Person auf YouTube, oder Social-Media-Postings mit der eigenen Stimme. Zeitgleich hat sich ein Markt für das Klonen von Stimmen entwickelt, z.B. für die Personalisierung von smarten Assistenten, Vorlesen von Hörbüchern oder für digitale Avatare (vicomtech 2023; Amezaga/Hajek 2022: 24). Auch Möglichkeiten zur Synchronisation («Dubbing») eines Sprechers in einem Video in einer anderen Sprache sind inzwischen einfach möglich (z.B. Masood u.a. 2022: 3986). Besonders wichtig sind hierbei die «Bewegung und das Aussehen des unteren Teils des Mundes und der ihn umgebenden Region» sowie die Synchronisation der Bewegungen mit Emotionen, welche sich jedoch bei unbekannten Gesichtern nicht generalisieren lassen (ebd.: 3986). So zeigen Prajwal u.a. (2020), dass die automatische Anpassung der Lippenbewegungen in einem Video mit der Tonspur möglich ist und somit ein englischsprachiger Schauspieler Deutsch, Spanisch und weitere Sprachen spricht. Anfangs führten diese Systeme noch zu erkennbaren und unrealistischen Verformungen des ganzen Gesichts, doch inzwischen gibt es Algorithmen, welche auf realistische Weise nur die Mundpartien anpassen (Sun u.a. 2022b: 1).

«Speech synthesis» (auch TTS genannt) kommt dann zum Einsatz, wenn keine vorhandene Stimme imitiert werden soll, sondern nur geschriebene Wörter durch eine Maschine ausgesprochen werden sollen. Traditionell gibt es hierzu zwei Ansätze: konkatenatives TTS und parametrisches TTS (Sciforce 2020):

- **Konkatenatives TTS** nutzt eine Datenbank mit Audioclips, welche dann zu ganzen Sätzen kombiniert werden. Das Ergebnis ist zwar verständlich, klingt aber typisch «roboterhaft». Wiedergabe von Emotionen oder weitere Intonationen sind damit kaum möglich (Sciforce 2020). Zudem muss für jede neue Stimme eine neue Datenbank mit Audioschnipseln erstellt werden.

- **Parametrisches TTS** nutzt ein Modell einer Stimme. Dieses Modell kann aus Aufnahmen einer Zielperson abgeleitet werden und kann auch den charakteristischen Klang und die Feinheiten der Aussprache einer Person erfassen. Ein Audioclip mit nur wenigen Minuten aufgezeichneter Sprache kann bereits ausreichen, um ein Modell der Stimme der Person zu erstellen. Normalerweise werden jedoch mehrere Stunden Audiomaterial als Trainingsgrundlage genutzt (Napolitano 2020: 5). Doch auch dann sind die Ergebnisse nicht perfekt. Aktuell eingesetzte Algorithmen nutzen nämlich nur einschichtige nicht lineare Transformationseinheiten, wie z.B. Hidden-Markov-Modelle (HMMs). Diese liefern gute Leistung bei Daten mit weniger komplizierten internen Strukturen – jedoch ist die Stimme sehr komplex (Ning u.a. 2019: 1 f.). Auf der Grundlage dieses Modells kann ein Computer neue Audioclips generieren, in denen ein beliebiger Text mit einem Klang ausgesprochen werden kann, der der Zielstimme sehr ähnlich ist. Ist bereits ein Modell vorhanden, lässt sich eine Stimme sehr einfach und schnell verändern (Napolitano 2020: 5). Seitdem das Training unter Rückgriff auf GANs erfolgt, konnte eine starke Verbesserung der Ergebnisse erzielt werden. Eine Bewertung dieser Ergebnisse ist jedoch schwierig, da die «Natürlichkeit» einer generierten Stimme nicht objektiv bewertet werden kann und zudem kein Vergleichsmaterial vorhanden ist (Mittag/Möller 2020: 1748). In der jährlich stattfindenden «Blizzard Challenge» versuchen verschiedene Teams daher ihre TTS-Systeme anhand der gleichen Datenbasis zu trainieren (SynSIG 2022). Die Ergebnisse werden dann durch Expertinnen und Experten in einem zeitaufwendigen «auditory listening test» verglichen und ein «naturalness mean opinion score» vergeben (Ling u.a. 2021: 3). Hierzu zählt unter anderem auch, wie viele Wörter falsch ausgegeben wurden («word error rate») (ebd. 2021: 7 ff.). Es gibt auch Versuche, die Natürlichkeit mittels KI zu bewerten, wie in Mittag/Möller (2020: 1751). Jedoch zeigen sich hier technische Schwierigkeiten, z.B. bei Audiodateien, welche per Telefon übertragen werden.

## 2.5. Technologien zum Generieren von inhaltlich authentischem Text

Mittels Technologien zur Textsynthese können automatisch Inhalte erstellt werden, die dem Sprach- oder Schreibstil der Zielperson ähneln (van Huijstee u.a. 2021: 13; euronews 2023). Um dies zu ermöglichen, nutzen die Technologien die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) (Vee-

rasamy/Pieterse 2022: 341). NLP ist eine Disziplin der Linguistik und Informatik und brachte seit 2011 sehr viele Anwendungsbereiche, wie z.B. die Spracherkennung in smarten Assistenten wie Amazon Alexa, maschinelles Übersetzen, oder die automatische Ergänzung von Suchbegriffen in Suchmaschinen hervor (Krohn u.a. 2020). Das Ziel von NLP ist es, Wissen und weitere Informationen (z.B. den Schreibstil) aus geschriebenem Text zu extrahieren – oder einfacher ausgedrückt: menschliche Sprache zu verstehen und zu interpretieren (van Huijstee u.a. 2021: 13). Während Verstehen und Interpretieren Aufgaben sind, die ein Mensch sehr gut durchführen kann, sind NLP-Systeme darin weniger gut, da diese darauf ausgelegt sind, eine grosse Menge an Text zu analysieren. Neben NLP gibt es Unterformen wie Natural Language Unterstanding (NLU), das reine Verstehen eines Textes (maschinelles Leseverständnis), und das Natural Language Generation (NLG), das sich mit der Konstruktion von Text befasst. Ein Eingangstext (ein Satz oder mehrere Sätze) wird in der lexikalischen Analyse in Wörter oder kleinere Einheiten («Token») zerlegt und deren Bedeutung ermittelt. In der nachfolgenden syntaktischen Analyse wird der Zusammenhang der «Token» im Satz ermittelt. Hierzu werden diese in eine standardisierte Struktur gebracht und deren Beziehung in einer hierarchischen Struktur dargestellt. Während der semantischen Analyse werden die syntaktischen Strukturen über mehrere hierarchische Ebenen und schliesslich über mehrere Sätze und den ganzen Eingangstext hinweg in Beziehung zueinander gesetzt, um deren jeweilige Bedeutung im Satz und im gesamten Text zu ermitteln. Zuletzt erfolgt die Transformation, mittels derer die Ausgabe erstellt wird. Je nach Anwendungsszenario (Vorgaben der Datenbank) kann dies eine Übersetzung, eine Grammatikkorrektur oder die Erstellung eines neuen Textes mit zugeordnetem Thema sein (Krohn u.a. 2020).

Einer der gängigsten Algorithmen, der in der NLP-Architektur eingesetzt wird, ist der «Transformer». Dieser ist in der Lage, aus Eingabetexten zu extrahieren, wie Wörter und Sätze in Beziehung zueinander stehen und daraus dann neue Texte zu generieren (Brown u.a. 2020: 3). Einer der fortschrittlichsten «Transformer»-Algorithmen ist der Generative Pre-trained Transformer 3 (GPT-3) (ebd.: 5). GPT-3 besteht aus 175 Milliarden Parametern und zeigte bereits bemerkenswerte Leistungen beim Übersetzen und Beantworten von Fragen (ebd.: 3; Brown u.a. 2020: 5). Aus den gelernten Informationen leitet der Algorithmus dann ein Modell der Sprachweise ab. Dieses dient dann wiederum zur Synthese neuer Reden (van Huijstee u.a. 2021: 13). Auch Nachrichtenartikel lassen sich so generieren, die nur schwer von menschengeschriebenen Artikeln zu unterscheiden sind, wie beispielsweise *The Guardian* im Jahr 2020 mit einem Nachrichtenartikel, der komplett durch GPT-3 erstellt wurde, demonstrierte (GPT-3 2020). Mit

ChatGPT in der Version 4.0 wurde ein mächtiges Tool geschaffen, welches nicht mehr nur den Schreibstil bekannter Autorinnen und Autoren nachahmen oder glaubwürdige Nachrichtenartikel verfassen kann. Mit etwas zusätzlicher Arbeit (z.B. Installieren einer Erweiterung namens «Human») kann es nämlich auch den Schreibstil von selbst eingegebenen Texten und damit von einer beliebigen Person imitieren (Mey 2023).

## 2.6. Technische Gegenmassnahmen

Die Gegenmassnahmen gegen Deepfakes lassen sich grundsätzlich in zwei verschiedene Kategorien aufteilen: Prävention und Erkennung (Mirsky/Lee 2022: 26).

### 2.6.1. Prävention

Wie in Kapitel 2.2.1. ausgeführt, ist das Katz-und-Maus-Spiel zwischen Erkennung und Generierung aktuell noch sehr volatil und es ist absehbar, dass Detektoren den Generatoren irgendwann unterlegen sein werden. Daher wird neben der Forschung an Detektoren auch an präventiven Massnahmen zur Verhinderung der Notwendigkeit der Detektion eines Deepfakes geforscht. Solche präventiven Ansätze werden in diesem Kapitel vorgestellt.

#### 2.6.1.1. Authentifizierungsverfahren zur Feststellung der Originalität oder Autorenschaft

Verschiedene technische Authentifizierungsverfahren sind in Entwicklung, um die Echtheit und Originalität von Bild-, Video- und Audioinhalten nachzuweisen.

Um die Originalität eines Bildinhalts nachvollziehen zu können, gibt es Vorschläge, die Distributed-Ledger-Technologie (DLT) einzusetzen. Mittels DLT kann ein Hashwert des originalen Bildes erstellt und gespeichert werden. Gibt es nachträgliche Veränderungen, werden sie im Hashwert sofort erkannt (Fraga-Lamas/Fernandez-Carames 2020). Allerdings lässt sich so nicht verhindern, dass eine Quelle das Material selbst fälscht, bevor es signiert (also in die Blockchain eingetragen) wird (BSI 2022a). Um dies zu verhindern und ggf. auch die Echtheit des dargestellten Bild- oder Audioinhalts zu gewährleisten, wird daran gearbeitet, eine Datei (etwa ein Video) direkt beim Aufnahmeprozess in der Kamera

mit einer digitalen Signatur zu versehen (ebd.). Zwar sind solche Ansätze technisch relativ einfach umsetzbar, dennoch werden sie aus mehreren Gründen kritisiert. Nachrichten und Geheimdiensten gäben sie die Möglichkeit, Urheber, die Quelle und Ort von Material auswerten zu können, wodurch Whistleblower, Menschenrechtsaktivisten etc. in Gefahr geraten können (Leetaru 2018). Zudem sind solche Authentifizierungsmechanismen mit einem grundsätzlichen Problem konfrontiert. Denn in den seltensten Fällen wird ein Event mit einer Kamera von Anfang bis Ende gefilmt. Bei einem Bombenanschlag z.B. wird meist erst nach der Explosion gefilmt (Leetaru 2018). Digitale Signaturen in Smartphone-Videos würden somit nichts weiter tun, als ein «falsches Video-Narrativ aufbauen, da die Frage nicht ist, ob das Material echt oder fake ist, sondern ob das Video die ganze Situation darstellt» (übersetzt aus ebd., 2018). Ebenso adressieren diese nicht das Problem, dass auch ein mittels Signatur belegbar originales Foto eine unechte Situation abbildet, die durch menschliches Hinzutun, z.B. durch Schauspielende, manipuliert wurde. Zudem gelang es Hackern in der Vergangenheit bereits, Signaturprozesse zu umgehen bzw. den digitalen Signaturprozess nachzukonstruieren (sog. «Reverse Engineering» und somit auch Deepfakes als echt zu deklarieren) (ebd.).

Die Vorteile der Blockchain sind stets spezifisch und vom Design und der Implementierung abhängig. So könnte die Speicherung eines Zeitstempels zwar beim Nachweis eines Dokuments helfen, aber gleichzeitig nicht Aufschluss über den echten Autor geben (Cheikosman u.a. 2021). Damit Endnutzende die Echtheit sicherstellen können, wären weitere Massnahmen wie Browser-Plugins nötig (Yazdinejad u.a. 2020). Ob und wie derartige Massnahmen in der Praxis durch die Nutzenden eingesetzt werden, ist jedoch fraglich.

Damit die Blockchain wirksam helfen kann, ist die internationale Zusammenarbeit von Techfirmen und Regierungen weltweit vorteilhaft (Cheikosman u.a. 2021). Microsoft hat in den letzten Jahren eine Kooperation mit BBC/CBC und der *New York Times* gestartet und das «Project Origin» ins Leben gerufen (Microsoft Innovation 2022). Dabei sollen fälschungssichere Metadaten[3] fest in einem Bild signiert werden und somit für jedes Bild ein eindeutiger Fingerabdruck generiert werden. Dabei kommen sog. Hashwerte zum Einsatz, einfach ausgedrückt handelt es sich dabei um Quersummen von Bilddaten. Sobald sich auch

[3] Metadaten sind Informationen über ein Bild, die in den sog. EXIF-Daten gespeichert werden. Diese sind primär unsichtbar, können aber mit Werkzeugen sichtbar gemacht werden. Typische Einträge dieser Bilder sind Kameramodell, Zeitpunkt der Aufnahme, GPS-Standort und Einstellungen der Kamera.

nur ein einziger Pixel oder Byte im Bild ändert, ist der Hashwert ein anderer. Zukünftig könnten dann beispielsweise Browser-Plugins aufzeigen, ob ein angezeigtes Bild tatsächlich echt ist. Über die Effektivität lässt sich aktuell noch nichts berichten. Zwar sind Hashwerte ein mathematisch eindeutiger Lösungsansatz und somit eine gute Herangehensweise, dennoch ist es mit genug Aufwand möglich, eine Datei mit gleichem Hashwert zu erstellen (Peyravian u.a. 1998). Die «Content Authenticity Initiative» ist ein von Adobe geführtes Konsortium bestehend aus vielen Zeitungen, Firmen, Kameraherstellern und Bildverwertern. Zu den Mitgliedern gehören unter anderem BBC, Twitter, dpa, getty images, Nikon, Qualcomm und Microsoft (Adobe 2022). Beide Initiativen sind unter dem Dach der «Coalition for Content Provenance and Authenticity» vereint (C2PA 2022). Diese möchte es Erstellern und Herausgebern von jedweden Inhalten (Bild, Video, Audio und Text) ermöglichen, zu belegen, woher der Inhalt kommt und dies einem Empfänger transparent zu machen. Ziel sei es, in den nächsten zwei bis fünf Jahren die Technologie in vielen Kamerasystemen einzuführen und bekannt zu machen. Neben dem Einsatz zur Erkennung von Deepfakes sehen die Mitglieder noch viele weitere Anwendungsgebiete, z.B. beim Einsenden von Schadensbildern bei Versicherungen oder Fotos von Hotels.

Zum Schutz vor Stimm-Deepfakes stellen Chu u.a. (2022: 2078) einen Algorithmus vor, welcher anhand eines Originalvideos die Lippenbewegungsmuster extrahiert und die Beziehung zwischen Gesichts- und Sprachmuster lernt. Ein Abgleich des Musters mit späteren Deepfake-Videos erlaube dann die Erkennung von Fälschungen. Allerdings ist es fraglich, ob diese Technologie erfolgreich sein kann, da andere Algorithmen diese Muster und Beziehungen ebenso lernen können und dann bei der Erstellung eines Deepfake-Videos anwenden.

#### 2.6.1.2. Störung von Generatoren

Generatoren können auch durch gezielte Gegenangriffe gestört werden. Hierbei werden die Trainingsdaten bewusst verändert (sog. gegnerisches [adversarial] maschinelles Lernen). Dies kann z.B. durch das Hinzufügen von einem bestimmten (für Menschen unsichtbaren) Muster geschehen (Sun u.a. 2022a: 1162) und dafür sorgen, dass Deepfake-Technologien ein richtiges Gesicht in einem Bild erkennen (Shawn u.a. 2020). Zudem wird diskutiert, dass den Generatoren in den nächsten Jahren das Trainingsmaterial ausgehen könnte und somit die Qualität fortan nicht mehr gesteigert werden könnte (Rivera 2023).

### 2.6.2. Erkennung

Die Erkennung von manipulierten Medieninhalten ist ein weites Forschungsgebiet mit vielen verschiedenen Methoden und basiert auf der Vorstellung: «Was KI kaputt gemacht hat, kann KI auch wieder reparieren» («What AI has broken can be fixed by AI as well») (Nguyen u.a. 2022: 12). Wie bereits in Kapitel 2.2.1. beschrieben, ist die Erstellung und Erkennung von Deepfakes ein Katz-und-Maus-Spiel zwischen Generatoren und Detektoren (Min-Max-Wettstreit). Aufgrund dieser technischen Grundlage ist grundsätzlich davon auszugehen, dass über einen langen Zeitraum der Generator-Teil des GAN gewinnen wird, weil dieser immer Rückmeldung über gefundene Fehler im Bild des Detektor-Teils bekommt und somit dem Detektor letztlich voraus sein wird. Zudem können selbst kleine Änderungen an den Generatoren dazu führen, dass Detektoren diese nicht mehr erkennen (Langguth u.a. 2021: 7).

Zum Zeitpunkt der Erstellung dieses Berichts zeigen GANs allerdings noch eine Reihe von Fehlerquellen – insbesondere dann, wenn der Detektor dem Generator nicht genügend Informationen über die gefundenen Fehler liefert (Google Developers 2020). Diese Probleme sind immer noch nicht gelöst und gegenwärtiger Stand der Forschung. Zur Überwindung der Probleme werden verschiedene Verfahren ausgetestet, z.B. modifizierte Min-Max-Funktionen oder der sog. «Wasserstein-Verlust», welcher erzwingt, dass der Detektor immer Feedback geben muss (Arjovsky/Bottou 2017: 12). Eine weitere Herausforderung ist der sog. «Mode-Collapse». Denn grundsätzlich versucht ein Generator zwar eine möglichst grosse Bandbreite an Ergebnissen zu produzieren. Wenn ein Generator jedoch eine besonders plausible Ausgabe erzeugt, kann er lernen, fortan nur diese Ausgabe zu erzeugen. In so einem Fall ist es die beste Strategie des Detektors, diese Ausgabe immer zurückzuweisen. Auch zur Umgehung dieses Problems wird auf den «Wasserstein-Verlust» zurückgegriffen. Innerhalb der nächsten Jahre sollen auch diese Probleme gelöst sein (Google Developers 2020).

Unterscheiden lässt sich die Erkennung in artefaktspezifische und ungezielte Ansätze (Mirsky/Lee 2022: 27 ff.), welche im Folgenden näher betrachtet werden.

#### *Artefaktspezifisch*

Deepfakes erzeugen oft Artefakte, die für Menschen unsichtbar sind, aber mittels forensischer Algorithmen und KI erkannt werden können. Mirsky/Lee (2022: 26 f.) unterscheiden hierzu sieben verschiedene Arten von Artefakten, welche genutzt werden können, um Deepfakes durch künstliche Intelligenz zu erken-

nen. Diese können räumlicher oder zeitlicher Natur sein und werden im Folgenden kurz vorgestellt.

**Räumliche Artefakte in Überblendungen:** Derartige Artefakte treten oft dann auf, wenn der generierte Inhalt wieder zurück in das Video überblendet wird. Um dies zu erkennen, haben Forschende Kantendetektoren entwickelt. Diese sollen die Grenzen von Inhalten in Videomaterial erkennen. Li u.a. (2020) haben hierzu ein neuronales Netzwerk aufgebaut, welches die Überblendungsgrenze eines Bildes vorhersagen und somit ein Video als echt oder falsch erkennen kann.

**Räumliche Artefakte in Umgebungen:** Es ist möglich, dass ein gefälschtes Bild im Zusammenhang mit dem Rest des Bildes (Umgebung) anormal ist. Dies könnten Rückstände bei Verformungen sein (Li/Lyu 2019), unrealistische Beleuchtung (Schattenwurf) und Unschärfen (Pu u.a. 2021: 983; Mirsky/Lee 2022: 26 f.).

**Räumliche Artefakte in den digitalen Daten:** GANs hinterlassen eindeutige Spuren (Fingerabdrücke), selbst wenn der Inhalt durch Kompression verändert wurde (Yu u.a. 2018). Koopman u.a. (2018) nutzen das natürliche Sensorrauschen einer Digitalkamera, um eingefügte Inhalte zu erkennen. Weiter zeigen genaue Analysen, dass GANs oft die Kopfhaltung, Augenzwinkern und andere physiologische Signale nicht korrekt generieren (Agarwal u.a. 2020: 1).

**Zeitliche Artefakte im menschlichen Verhalten:** Dank grosser Datenmengen können Verhaltensweisen von bekannten Personen analysiert werden. Agarwal u.a. (2019) untersuchten auf diese Weise Videos auf Anomalien in der Verhaltensweise der abgebildeten Person. Dies ist – mit entsprechend abnehmender Qualität – auch möglich, wenn eine geringere Menge an Referenzmaterial zur Verfügung steht (Mittal u.a. 2020).

**Zeitliche Artefakte in der Physiologie:** Neben Bewegungen gibt es menschliche physiologische Signale, die in einem Deepfake-Video analysiert werden können (Mirsky/Lee 2022: 27). Ciftci u.a. (2020) nutzen zum Beispiel den Puls und somit die Blutbewegungen in den Gefässen und Änderungen der Hautfarbe, um Fälschungen zu identifizieren.

**Zeitliche Artefakte in der Synchronisation:** Unstimmigkeiten im zeitlichen Ablauf eines Videos sind ebenfalls analysierbare Merkmale. So weisen Deepfake-Videos minimale Abweichungen zwischen Mundform und den gesprochenen Wörtern (Phoneme). Besonders aufschlussreich seien hier Phoneme, bei denen der Mund vollständig geschlossen ist (z.B. beim Buchstaben «M») (Agarwal u.a. 2020). Auf diese Art der Erkennung fokussieren sich Algorithmen, welche Fälschungen anhand der Stimme erkennen (Zhou/Lim 2021: 14780; Korshunov/

Marcel 2018; Masood u.a. 2022: 3988 ff.). So stellen u.a. Shahzad u.a. (2022) mit «Lip Sync Matter» eine Erkennungsmethode vor, welche die Diskrepanz zwischen der aus dem Video extrahierten Lippensequenz und der synthetischen Lippensequenz aus dem Audiomodell anhand von semantischen Merkmalen erkennt.

**Zeitliche Artefakte in der Kohärenz aufeinanderfolgender Bilder:** Deepfake-Videos produzieren oft Flackern und Flimmern in aufeinanderfolgenden Bild-Frames (Mirsky/Lee 2022: 9). Diese können erkannt werden, wie beispielsweise Sabir u.a. (2019) herausfanden.

*Ungezielte Ansätze*

Anstatt dass Erkennungsalgorithmen vorab auf ein bestimmtes Artefakt fokussiert werden (supervised learning), können neuronale Netze auch selbst entscheiden, welche Merkmale eines Videos analysiert werden sollen (unsupervised learning) (Mirsky/Lee 2022: 27). Besonders relevant sind in diesem Fall die sog. gefalteten neuronalen Netze (convolutional neural networks, CNN). Diese bestehen aus mehreren Faltungsschichten (Ebenen), wobei die einzelnen Neuronen zwei- oder dreidimensional miteinander verbunden sind. Diese Technologie bietet sich vor allem bei der Verarbeitung von Video- und Tonmaterial an. Hierzu gibt es zwei Ansätze: Klassifizierung und Anomalieerkennung.

**Klassifizierung:** CNN schneiden bei der Erkennung von komprimiertem Videomaterial besser ab als die bisher vorgestellten forensischen Werkzeuge (Nguyen u.a. 2019: 2307). Hierzu trainierten unter anderem Hsu u.a. (2020) ein CNN mit kontrastierendem Trainingsmaterial aus echten und gefälschten Bildern. Ein Nachteil ist jedoch, dass dadurch nur Fakes erkannt werden können, auf die das CNN trainiert wurde. Dieses Problem wurde mittels Hierarchical Memory Network (HMN) adressiert. HMN berücksichtigen den Inhalt des Gesichts und zuvor gesehener Gesichter (Fernando u.a. 2019). Zudem kommen mehrere CNN gleichzeitig zum Einsatz (Rana/Sung 2020) und 3D-CNN können mehrere Frames (Bilder) gleichzeitig betrachten (Lima u.a. 2020). Dennoch sind Klassifikationen mittels CNN anfällig, da ein Angreifer die Erkennung selbst durch maschinelles Lernen umgehen kann (Mirsky/Lee 2022: 28).

**Anomalieerkennung:** Hierbei werden die Modelle auf echte Daten trainiert und erkennen dann Anomalien während des Einsatzes auf Deepfakes (Mirsky/Lee 2022: 28). Dies hat den Vorteil, dass vorab keinerlei Angaben gemacht werden müssen, welche Art und Form der Fake haben könnte, sodass auch bisher «unbekannte» Artefakte erkannt werden können.

Da Menschen keine Chance mehr haben, Deepfakes zu erkennen, sind die o.g. Methoden zur Erkennung aktuell das wichtigste Werkzeug. Da die aktuellen Deepfakes Bild für Bild erstellt werden, können diese Werkzeuge gegenwärtig dafür sorgen, dass die Detektoren den Generatoren einen Schritt voraus sind (Agarwal u.a. 2020: 1). Dass dies jedoch nicht mehr allzu lange der Fall sein wird, zeigen Untersuchungen, welche den Detektoren eine abnehmende Performance zuschreiben (Mirsky/Lee 2022: 31).

*Software zur Erkennung von Deepfakes*

Neben der Erkennung von Deepfake-Inhalten durch blosse Betrachtung ohne Zuhilfenahme technischer Hilfsmittel, hat sich ein Markt für technische Detektionsanwendungen als konsumentenorientierte Dienstleistung entwickelt. Auch in wissenschaftlichen Publikationen werden Produkte vorgestellt und deren Effektivität und Anwendung diskutiert (Rana u.a. 2022; Mirsky/Lee 2022).

Grundsätzlich spezialisieren sich Forschende beim Thema Detektion auf bestimmte Video- und Audioartefakte, wie beispielsweise Zwinkern (Jung et al. 2020) oder andere (z.B. visuell-zeitliche) Merkmale, die als schwer manipulierbar bzw. synthetisierbar gelten (Montserrat et al. 2020). Oder sie fokussieren sich auf bestimmte Anwendungsfälle, wie Mehta et al. (2021), auf Livebilder bei Videokonferenzen mit Zoom und Skype. Diese werden dann anhand einiger Datensätze von Deepfakes getestet. Diese Datensätze bestehen aus 49 bis 202 000 Bildern und/oder Videodateien. Rana et al. (2022, S. 25501) listen diese auf und zeigen, wie oft diese in Studien bisher verwendet wurden. Demnach wird der «FaceForensics++»-Datensatz mit Abstand am meisten in wissenschaftlichen Studien eingesetzt (Rossler et al. 2019). Er beinhaltet 1000 Videos mit mehr als 1.8 Millionen Bildern von generierten Gesichtern. Aufgrund der Vielfalt der genutzten Datensätze gestaltet sich eine Generalisierung von Aussagen über die Treffergenauigkeit von Detektoren dennoch als schwierig (Nguyen u.a. 2022: 12).

In den letzten Jahren wurde eine Vielzahl von Detektoren zur Erkennung von manipulierten und synthetisierten Text-, Audio- und Bildinhalten entwickelt und vermarktet, z.B. *AI Classifier, GPT-0* oder *Originality.AI*. Mit «Dessa» gibt es seit 2019 ein Open-Source-Programm zur Erkennung von Deepfake-Audios (Dessa 2019). Auch grosse Unternehmen wie Microsoft und Adobe arbeiten an grossflächigen Diensten zur Erkennung von Deepfake-Inhalten (TechRepublic 2021). Die Anwendungen sind allerdings mit dem Problem dürftiger Erkennungsraten konfrontiert (Tangermann 2023). Obwohl einige Methoden hohe

Erfolgsraten haben, liegt dies oft daran, dass Daten manuell vorbereitet werden müssen («The model achieved a 98 % success rate in detecting tasks, but the data needed to be pre-processed manually to extract the relevant features») (Almutairi/Elgibreen 2022: 4). Trotz der jahrelangen Diskussion und Entwicklung von Detektoren sind bislang nur wenige Produkte auf dem Markt, die frei nutzbar sind. Auch der bereits 2020 angekündigte Microsoft Video Authenticator ist noch nicht veröffentlicht und wird in absehbarer Zeit wohl auch nicht mehr den Markt betreten. Zwischen den in Detektionstechnologien gesteckten Hoffnungen und der realen Erkennungsleistung klafft ausserdem eine grosse Lücke (Yazdinejad u.a. 2020: 5). Als möglicher Lösungsweg wird die Bewertung von Bildern gemeinsam durch KI und Mensch diskutiert (Chen u.a. 2019).

Im Jahr 2019 wurden drei Benchmark-Datenbanken von Google, Facebook und Celeb-DF veröffentlicht. Die darauf angesetzten Detektoren erreichten jedoch nur schlechte Erkennungswerte (AUC = 0,86; 0,76 und 0,66[4]) (Mirsky/Lee 2022: 31). Zudem liegt die Fehlerrate (false-positive) bei 0,001 (0,1 %), was nach einem geringen Wert klingt, jedoch angesichts der vielen Millionen Bilder, welche täglich ins Internet gestellt werden, sehr viele als falsch erkannte Bilder bedeutet (ebd.: 31).

Technisch ergeben sich weitere Herausforderungen für die Detektoren:

- Um einen Detektor zu umgehen, der sich auf die Erkennung von Artefakten spezialisiert hat, muss ein Angreifer nur eine einzige Schwachstelle finden, die dann zur Generierung neuer, darauf abgestimmter Bilder verwendet werden kann (Mirsky/Lee 2022: 31; Wang u.a. 2019).
- Detektoren, die Fakes anhand von Unschärfe erkennen, können durch GANs, die bewusst Rauschen oder Schärfe generieren, umgangen werden (Jalalifar u.a. 2018).
- Detektoren, die Ränder von eingefügten Inhalten in Bildern erkennen, können durch GANs, die ganze Vollbilder erstellen, umgangen werden (Kim u.a. 2018).

Auch die beschriebenen Klassifikatoren lassen sich austricksen, indem bewusst Störungen und Rauschen durch eigene KI-Methoden ins Bild eingebracht werden (adversarial machine learning). Dies sei inzwischen auch ohne Kenntnis des Klassifikators oder seiner Trainingsdaten möglich (Carlini/Farid 2020).

---

[4] Die ermittelte Fläche unter der Kurve (Area Under the Curve, AUC) ist das Mass für die Fähigkeit eines Klassifikators, zwischen Klassen zu unterscheiden. Das Maximum wäre somit 1,0.

Seit den jüngsten Fortschritten bei KI-generierten Texten sind inzwischen auch Detektoren für KI-generierte Texte entwickelt worden. OpenAI hatte mit dem *AI Classifier* im Januar 2023 eine Software vorgestellt, welche erkennen sollte, ob es sich um generierten Text oder menschlichen Text handele (OpenAI 2023). Im Juli 2023 wurde diese Software jedoch wieder zurückgezogen (Epstein-Gross 2023). Dies ist laut OpenAI der schlechten Erkennungsrate von 26 % sowie der false positives-Rate von 9 %, also der falschen Erkennung eines Textes als KI-generiert, obwohl er von einem Menschen stammt, geschuldet (OpenAI 2023; Epstein-Gross 2023). Jedoch würden die Forschenden nun an «wirksameren Verfahren zur Herkunftssicherung von Texten» (wörtliche Übersetzung aus OpenAI 2023) arbeiten und die Erkennung auf audiovisuelles Material erweitern. Auch im Falle der Software «Turnitin» – eine Software, welche die Erkennung von Plagiaten an Universitäten verbessern soll (Turnitin 2023) – wurde der Einsatz aufgrund der schlechten Erkennungsraten schnell zurückgezogen (Epstein-Gross 2023). Dies führt u.a. dazu, dass einige US-amerikanische Universitäten wieder auf Handschrift setzen – auch wenn die handschriftliche Abschrift KI-generierter Texte trotzdem möglich ist (Hart/Mok 2023). Die verbesserungsbedürftigen Trefferraten der Detektionsanwendungen führen mitunter auch zu amüsanten Ergebnissen, wie beispielsweise, dass die Verfassung der USA als KI-generierter Text detektiert wurde (Edwards 2023). Weniger amüsant sind hingegen Fehlerkennungen, die zur Diskriminierung führen können. Denn Texte, welche von nicht englischen Muttersprachlern geschrieben wurden, würden häufiger fälschlicherweise als KI-generiert erkannt, wodurch diese Texte bzw. Autorinnen und Autoren benachteiligt werden können (ebd.).

Abschliessend sei erwähnt, dass künftig architektonische Probleme der Generatoren, sofern sie nicht überwunden werden, Detektoren Vorteile bringen können. Denn einige Forschende prognostizieren, dass es KIs bald an Textdaten für das weitere Training mangeln wird. Dies schliesse Bücher, Nachrichtenartikel, wissenschaftliche Artikel und Wikipedia-Einträge ein. Ausserdem wird davon ausgegangen, dass ChatGPT und anderen Text-Generatoren bis 2026 hochwertige Trainingsdaten ausgehen werden (Rivera 2023), was die weitere Entwicklung von Generatoren verlangsamen werde. Konnten KIs bislang mit der Summe der in den vergangenen Jahrzehnten hauptsächlich von Menschen generierten Daten trainiert werden, werden KIs künftig einerseits auf jeweils neu entstehende Inhalte beschränkt sein und andererseits unter dem Problem leiden, dass ein zunehmend grosser Teil dieser neu entstehenden Inhalte selbst KI-generiert sein wird. Falls es so kommt, würde die weitere Entwicklung der Generatoren derart verlangsamt, dass die Entwicklung von Detektoren aufholen könnte.

Erwähnt sei auch, dass bereits an der Überwindung solcher Probleme geforscht wird. OpenAI und andere Anbieter von Generatoren sind auf der Suche nach neuen Datenquellen. ChatGPT nutzt bereits Daten, die aus Social-Media-Plattformen gewonnen wurden (Botpress Blog 2023). Allerdings bringen solche Lösungsstrategien wieder neue Probleme mit sich: Weil die auf diese Weise verarbeiteten Daten in der Regel zumindest im Widerspruch zu europäischen Datenschutzgesetzen stehen, ist die künftige Nutzung von zumindest europäischen Social-Media-Daten nämlich fraglich (Rivera 2023, vgl. auch Kapitel 3). Auf der Suche nach neuen Datenquellen konnte ein südkoreanisches Forscherteam mit DarkBERT eine ChatGPT-Alternative entwickeln, welche Daten aus dem Darknet nutzt (Jin u.a. 2023).

## 2.7. Untersuchung von Deepfake-Detektoren

In dieser Studie wurden kostenfreie Detektoren mit echtem und KI-generiertem Videomaterial ausgetestet. Unser Vorgehen ist in Abbildung 8 zu sehen. Im ersten Schritt wurde Videomaterial ausgewählt. Hierbei wurde keiner der Datensätze genutzt, die in Rana u.a. (2022: 25501) beschrieben werden. Angesichts der Problematik, dass bisherige Tests mit diesen Datensätzen schwer vergleichbare Ergebnisse produzierten (ebd.: 25501), wurde im Rahmen unseres Tests angestrebt, die Erkennungsleistung anhand von einfachen und bekannten Beispielen zu bewerten. Hierfür wurden zehn Videos politischer Persönlichkeiten ausgewählt (via YouTube, nach Anzahl der Ansichten), über die viel Bildmaterial im Internet vorhanden ist, sodass die Wahrscheinlichkeit, dass Detektoren auf deren Erkennung trainiert sind, erhöht ist. Pro Persönlichkeit wurde je ein Original und ein Deepfake gegenübergestellt (siehe Tabelle 1). Hierbei handelte es sich um Angela Merkel (frühere Bundeskanzlerin der BRD), Barack Obama (früherer Präsident der Vereinigten Staaten), Elisabeth II. (ehemalige Königin des Vereinigten Königreichs), Donald Trump (früherer Präsident der Vereinigten Staaten) und Volodymyr Selenskyjy (aktueller Präsident der Ukraine). Zudem nutzten wir ein selbst erstelltes Video, von dem wir mit Sicherheit wissen, dass es sich um ein Original handelt. Die Länge der Videodateien variierte zwischen 15 Sekunden und 7:42 Minuten. Das selbst erstellte Video diente dem Zweck zu sehen, wie die Detektoren mit potenziell unbekanntem Material umgehen.

**Tabelle 1:** Übersicht der getesteten Videos

| Video | Original/ Deepfake | Länge der Videos | URL |
|---|---|---|---|
| **Merkel_1** | Deepfake | 0:15 | https://www.youtube.com/watch?v=9-i75FmshZw |
| **Merkel_2** | Original | 7:30 | https://www.youtube.com/watch?v=N_7eYLpynOs |
| **Obama_1** | Deepfake | 1:12 | https://www.youtube.com/watch?v=cQ54GDm1eL0 |
| **Obama_2** | Original | 4:30 | https://www.youtube.com/watch?v=k0jJL_YFyIU |
| **Queen_1** | Deepfake | 3:45 | https://www.youtube.com/watch?v=yyFQPjIrnQE |
| **Queen_2** | Original | 7:42 | https://www.youtube.com/watch?v=KgvZnxNAThM |
| **Trump_1** | Deepfake | 0:42 | https://www.youtube.com/watch?v=U6XLg7398XM |
| **Trump_2** | Original | 6:41 | https://www.youtube.com/watch?v=RxTZxoZE4Yc |
| **Selensky-jy_1** | Deepfake | 1:12 | https://www.youtube.com/watch?v=X17yrEV5sl4 |
| **Selensky-jy_2** | Original | 1:25 | https://www.youtube.com/watch?v=prfaWHQoxVg |
| **Forum Privatheit** | Original | 5:39 | https://www.forum-privatheit.de/wp-content/uploads/Jahreskonferenz2022_Videos/Tag2_01_Er%C3%B6ffnung_klein.mp4 |

Anschliessend wurden kostenlose Detektoren recherchiert, um im nächsten Schritt, falls nötig, Zugang zu erhalten. Der vierte Schritt war der eigentliche Test der Programme. Hierfür wurden die Test-Videodateien entweder über die entsprechenden Websites als Datei hochgeladen oder die zugrunde liegende URL angegeben. Im letzten Schritt wurden die Ergebnisse der verschiedenen getesteten Detektoren gesammelt und dokumentiert.

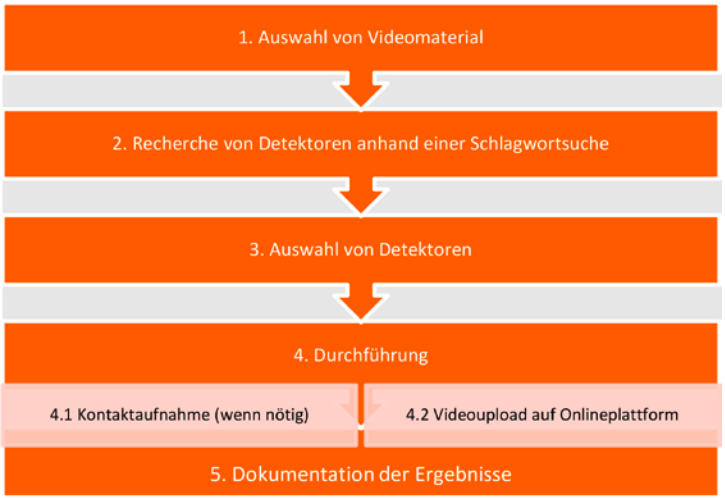


**Abbildung 8:** Ablauf der Untersuchung von Detektoren

### 2.7.1. Recherche von Detektoren und Kontaktaufnahme

Mittels einer Schlagwortsuche bei einschlägigen Suchmaschinen (Google, Bing und Komo.ai) wurden die Detektoren recherchiert. Im Rahmen dieses Prozesses konnten einige Anbieter identifiziert werden, die entweder bereits Produkte im Bereich der Deepfake-Erkennung anbieten bzw. momentan an entsprechenden Produkten arbeiten. Basierend auf dieser Liste von Anbietern wurden jene Anbieter isoliert, die ein bereits verfügbares Produkt für die isolierte Anwendung auf Videos anbieten. Stand Ende Mai 2023 umfasste dies sechs Anbieter (siehe Tabelle 2).

**Tabelle 2:** Anbieter von Detektoren

| # | Anbieter | URL |
|---|---|---|
| 1 | Deepware | https://deepware.ai/ |
| 2 | Reality Defender | https://realitydefender.com/ |
| 3 | Sensity AI | https://sensity.ai/ |
| 4 | DuckDuckGoose | https://www.duckduckgoose.ai/ |
| 5 | DeepFake-o-meter | https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/ |
| 6 | MeVer | https://mever.gr/https://mever.gr/ |

Da die Anbieter (mit Ausnahme von Deepware und des DeepFake-o-meter der University at Buffalo) den Zugang beschränken, wurden diese mit der Bitte um Zugang kontaktiert. MeVer gewährte Zugang am 06.06.2023. Die übrigen Anbieter wurden im Zeitraum zwischen dem 11.05.2023 und dem 26.05.2023 mehrfach und über unterschiedliche Wege kontaktiert, reagierten jedoch bis zum Zeitpunkt des Verfassens (September 2023) auf keine unserer Anfragen. Da DeepFake-o-meter eine dauerhafte Fehlermeldung aufgrund einer Serverwartung hatte, wurde der Anbieter am 01.06.2023 und 06.06.2023 über das entsprechende Kontaktformular kontaktiert. Allerdings erhielten wir keine Rückmeldung und der Fehler bestand zum Testzeitpunkt weiterhin. Daher konnte der Detektorentest letztlich mit zwei (Deepware und MeVer) der sechs Programme durchgeführt werden.

### 2.7.2. Resultate

Im Folgenden werden die detaillierten Ergebnisse von zwei Anbietern vorgestellt. Eine Übersicht der Ergebnisse finden Sie im Onlineappendix.[5]

*Deepware*

Generell bietet Deepware eine einfache und öffentlich zugängliche (also ohne die bei anderen Detektoren erforderliche Vorabgenehmigung der Nutzung) Möglichkeit, die Wahrscheinlichkeit für das Vorliegen von KI-basierten Manipulationen in Videos zu ermitteln. Die genauen technischen Grundlagen des Vorgehens können auf der entsprechenden GitHub-Seite (Deepware 2023a) gefunden werden. Vereinfacht ausgedrückt, durchläuft das getestete Video bis zu fünf Modelle bzw. Algorithmen, die durch Überwachung von systemweiter Veränderung der Verteilung der Hardware-Performance-Counter (HPC) Daten (= Geschwindigkeit bestimmter Vorgänge in einem Prozessor) nach Anzeichen von Manipulationen sucht. So können beispielsweise künstliche Veränderungen der im Video enthaltenen Gesichter oder anderer Elemente identifiziert werden. Hierbei bestimmt das Programm selbstständig, welche Modelle durchlaufen werden. Basierend auf dieser Analyse berechnet das Programm für jedes Modell die Wahrscheinlichkeit etwaiger KI-Manipulationen und klassifiziert das Video als eine der drei Ergebnisklassen:

[5] https://zenodo.org/records/10847902.

- «Deepfake detected» – Ein Deepfake wurde erkannt.
- «Suspicious» – Es liegt ein berechtigter Verdacht vor, dass ein Deepfake vorliegt.
- «No deepface detected» – Es wurde kein Deepfake erkannt.

Auffällig bei den Videos von **Angela Merkel** war, dass nur eines der fünf Modelle das Deepfake-Video als eindeutig manipuliert klassifizierte. Zwei Modelle klassifizierten das verfälschte Video als verdächtig. Die verbleibenden zwei Modelle, darunter der Algorithmus von Deepware, konnten keine Verfälschungen nachweisen. Das Ergebnis des Deepware-Algorithmus deutete dabei jedoch darauf hin, dass in dem kurzen und schlecht aufgelösten Video kein Gesicht durch den Algorithmus erkannt wurde (Deepware 2023b). Das Originalvideo, sprich, das nicht verfälschte Video, wurde von vier Modellen getestet und dabei von drei Modellen als unbearbeitet erkannt. Gleichzeitig wurde das Video durch den Algorithmus von Deepware mit einer 53%igen Manipulationswahrscheinlichkeit als verdächtig eingestuft. Dies ist nur etwas mehr als Zufall (also 50 %). In Angesicht dieser divergenten Ergebnisse kann das Gesamtergebnis nur eingeschränkt als korrekt betrachtet werden.

Beim Video von **Barack Obama** wurde die Manipulation lediglich durch einen der Algorithmen erkannt. Weitere drei durchlaufene Modelle stuften es als nicht manipuliert ein, wobei die angegebene Wahrscheinlichkeit für eine Manipulation dieser Modelle zwischen 0 und 28 % betrug. Das Originalvideo wurde nur durch den Deepware-Algorithmus getestet, welcher das Video fälschlicherweise als verdächtig einstufte. Die Wahrscheinlichkeit einer Manipulation wurde dabei mit 77 % beziffert.

Die Testergebnisse der Videos der **Weihnachtsansprache von Elisabeth II.** ähnelten den übrigen Ergebnissen: Während eines der Modelle den Fake als einen solchen erkannte, klassifizierten die anderen vier durchlaufenen Modelle das Video als unbedenklich. Die Wahrscheinlichkeiten lagen bei den Fehlklassifizierungen zwischen 3 und 20 %. Darüber hinaus wurde das Original erneut nur vom Deepware-Algorithmus geprüft, welcher das Video mit einer Wahrscheinlichkeit einer Manipulation von 94 % als Deepfake fehlklassifizierte.

Der Scan des Deepfake-Videos von **Donald Trump** schlug trotz mehrerer Versuche fehl, weshalb hierfür keine Ergebnisse vorliegen. Im Gegensatz dazu wurde das Original von allen vier durchlaufenen Modellen korrekt eingestuft. Hierbei lag die Manipulationswahrscheinlichkeit aller Modelle unter 4 %.

Das 2023 bekannt gewordene Deepfake-Video des ukrainischen Präsidenten **Volodymyr Selenskyjy** wurde abermals nur von einer der fünf Methoden als solches identifiziert. Die anderen vier Algorithmen, darunter erneut der Deepware-Algorithmus, stuften das Video mit Wahrscheinlichkeiten unterhalb von 3 % fälschlicherweise als unbedenklich ein. Das Original wurde mit Manipulationswahrscheinlichkeiten zwischen 0 und 11 % von allen vier genutzten Methoden korrekt als Video ohne Manipulationen eingestuft.

Das von uns selbst erstellte Video des **Forum Privatheit** wurde von der Hälfte der Algorithmen von Deepware als Fake klassifiziert. Der Deepware-Algorithmus gibt eine Erkennung von 30 % Wahrscheinlichkeit an. Leider liefert keiner der Algorithmen eine Information darüber, woran er dies festgemacht hat. Dass genau die Hälfte der Algorithmen ein Fake erkannten, ist für Nutzende bei der Entscheidungsfindung nicht hilfreich, da die 50 % genau dem Zufall entsprechen.

*MeVer*

MeVer entwickelte ein Programm, welches mithilfe eines multimethodischen Ansatzes ein einziges Wahrscheinlichkeitsergebnis für mögliche, KI-basierte Manipulationen des getesteten Videos bzw. Bildes ermittelt. Dabei werden nach einer Vorverarbeitung die in dem Testmedium enthaltenen Gesichter zuerst extrahiert und einem Gesamtsystem aus fünf Deepfake-Erkennungsmethoden zugeführt. Zuletzt werden die Methodenergebnisse aggregiert, um den gesamten Wahrscheinlichkeitsindikator zu ermitteln. Die genaue technische Implementation ist in Baxevanakis u.a. (2022) erläutert.

Bei dem ersten getesteten Video, dem Deepfake von **Angela Merkel**, schlug der Scan fehl. Hierdurch sind für dieses Video keine Ergebnisse vorhanden. Das Fehlschlagen des Scans ist dabei sehr wahrscheinlich in den Eigenschaften des Videos begründet. Aufgrund der schlechteren Auflösung, der Ganzkörperaufnahme und der kurzen Videolänge ist es naheliegend, dass das Programm kein Gesicht extrahieren und somit auch nicht untersuchen konnte. Dem Originalvideo von Angela Merkel ordnete der Algorithmus fälschlicherweise eine hohe Wahrscheinlichkeit von 84 % für Manipulationen zu. Bei Betrachtung der genaueren Ergebnisse zeigte sich, dass einige Keyframes mit einem Rotstich sehr hohe Scorings erhielten und hierdurch für den hohen Wahrscheinlichkeitswert verantwortlich sind.

Dem Deepfake-Video von **Barack Obama** wurde eine sehr geringe Manipulationswahrscheinlichkeit von lediglich 1 % zugeordnet und dieses somit fehlklassifiziert. Bei genauerer Betrachtung der Ergebnisse zeigte sich auch, dass der Algorithmus für einige Keyframes eine hohe individuelle Manipulationswahr-

scheinlichkeit ermittelte. Für das Originalvideo von Barack Obama berechnete der multimethodische Ansatz einen Gesamtscore von 13 %. Somit wurde das Video korrekt eingeordnet. Es zeigte sich jedoch erneut, dass einzelne Keyframes mit Rotstich die Gesamtwahrscheinlichkeit negativ beeinflussten.

Für das verfälschte und originale Video der **Weihnachtsansprache von Elisabeth II.** ermittelte der Ansatz von MeVer eine Gesamtwahrscheinlichkeit von 8 bzw. 11 %. Somit wurde das manipulierte Video fehlklassifiziert, während das Original korrekt bewertet wurde. Abermals ermittelte das Programm bei dem Deepfake unterschiedliche Wahrscheinlichkeiten für unterschiedliche Videoabschnitte. So wurden einigen Abschnitten Manipulationswahrscheinlichkeiten von bis zu 84 % zugeordnet. Gleichzeitig erhöhten einzelne Keyframes mit Rotstich den Gesamtscore des Originals.

Die Ergebnisse für die Videos von **Donald Trump** ähnelten jenen von Elisabeth II. So wurde das manipulierte Video mit einem Gesamtwert von 15 % ebenfalls fehlklassifiziert. Auch wenn erneut unterschiedliche Ergebnisse für unterschiedlich Abschnitte dieses Videos ermittelt wurden, fielen diese Wahrscheinlichkeiten mit bis zu 35 % vergleichsweise gering aus. Gleichzeitig erhöhten einzelne Keyframes des Originals dessen Manipulationswahrscheinlichkeit auf 13 %. Nichtsdestotrotz schätzte der Ansatz das Originalvideo korrekt ein.

Der Deepfake von **Volodymyr Selenskyjy** wurde mit einer Manipulationswahrscheinlichkeit von 84 % von MeVer korrekt erkannt. Im Kontrast hierzu wurde dem Original ein Gesamtwert von 44 % zugeordnet. Dabei beeinflussten einige Keyframes mit Rotstich erneut diese Gesamteinschätzung. Somit ist die Einschätzung des originalen Mediums für den Fall des ukrainischen Präsidenten unklar.

Im unbekannteren Video des **Forum Privatheit** erkannte MeVer mehrere Gesichter korrekt und gab den «Fake Score» mit 10 % an. Da es sich um ein echtes Video handelt, ist dieser Wert realistisch.

### 2.7.3. Fazit zu Deepfake-Detektoren

Deepfake-Detektoren gelten als ein wichtiger Baustein zur Erkennung von Deepfakes. Insgesamt sind die Ergebnisse unseres Detektoren-Tests in der Praxis jedoch unbefriedigend. Die getesteten Detektoren gaben sowohl falsch-negative als auch falsch-positive Ergebnisse aus. Problematisch war auch, dass Originalvideos fälschlicherweise als Deepfakes deklariert wurden, sodass durch die

Nutzung von Detektoren nicht nur Deepfakes nicht als solche erkannt werden können, sondern auch die Vertrauenswürdigkeit echter Inhalte untergraben werden könnte. Zudem bewegten sich die angegebenen Wahrscheinlichkeitswerte häufig im 50%igen Bereich, was angesichts der Hoffnung, Deepfake-Detektoren zur schnellen Begutachtung verdächtiger Inhalte zu nutzen, ebenfalls unbefriedigend ist. Ein Muster, das sich zeigte, ist, dass bekannte Deepfake-Videos erkannt werden. Diese sind z.B. auf YouTube weit verbreitet und es ist anzunehmen, dass die Detektoren hier hauptsächlich ein Vergleich des Videos mit dem bekannten, als Fake detektierten, Video durchführen und keine Analyse der eigentlichen Inhalte. Ein möglicher Anhaltspunkt hierfür ist, dass die Methode «Analyst» (siehe Onlineappendix[6]) die korrekteste ist. Diese Methode ist laut Beschreibung von einem Menschen durchgeführt worden – und schmälert die Qualität der softwarebasierten Ergebnisse weiter.

Unsere Tests mit den zwei nutzbaren Detektoren als auch unsere vorherige Diskussion des Stands der Literatur zu Detektoren zeigen somit, dass die Hoffnungen, die seit geraumer Zeit in Detektoren gesteckt werden, aktuell nicht erfüllt werden. Auch der Umstand, dass die Detektoren von lediglich zwei der sechs recherchierten Anbietern tatsächlich nutzbar waren, zeugt davon, wie wenig nützlich Detektoren zum Testzeitpunkt waren. Allein dass mehrere Detektoren nur nach vorheriger Genehmigung nutzbar gewesen sind, stellt ein Hindernis für den praktischen Einsatz solcher Werkzeuge dar: Schliesslich geht es beim Thema Deepfakes häufig um die Geschwindigkeit, mit der irreführende Inhalte entlarvt werden. Ein Genehmigungsprozess, der Wochen benötigt, ist dabei kontraproduktiv. Die Ergebnisse legen daher nahe, dass sich die Gesellschaft nicht auf die Ergebnisse aktueller Detektoren verlassen sollte.

## 2.8. Bibliometrische Auswertung wissenschaftlicher Publikationen

Zur Erfassung des wissenschaftlichen Publikationsaufkommens sowie der Positionierung der Schweiz im internationalen Vergleich wurde eine bibliometrische Auswertung durchgeführt.

[6] https://zenodo.org/records/10847902.

### 2.8.1. Methodik

Zur Auswertung wurde auf die Publikationsdatenbank *Scopus* mittels Structured Query Language (SQL) zugegriffen. Von Interesse waren die Publikationen zu den Themen Deepfakes und der Synthese von Video- und Tonmaterial der zehn publikationsstärksten Länder und zusätzlich der Schweiz (CH), Deutschland (DE) und EU (EU+CH+GB). Daraus ergeben sich folgende Länder: Australien (AU), Kanada (CA), Schweiz (CH), China (CN), Deutschland (DE), Frankreich (FR), Grossbritannien (GB), Indien (IN), Italien (IT), Japan (JP) und USA (US). Es wurden ausschliesslich deutsch- und englischsprachige Publikationen untersucht.

Für die Suche wurden mittels SQL-Zugang zur Scopus-Datenbank präzise Suchanfragen zur Auswahl von relevanten Publikationen erstellt. Die Suche fand im September 2022 statt, wobei die Datenbank alle Publikationen bis Kalenderwoche 17 des Jahres 2022 enthielt. Die Suche wurde auf die Schlagworte (Author_keywords), den Abstract und den Titel der Publikationen angewandt. Es wurden zwei verschiedene Suchabfragen durchgeführt. Die erste mit den Begriffen: *Deepfake%* | *Deep fake%* | *Deep-fake%.*[7]

Für eine zweite Suche wurden weitere Schlagworte aus vier einschlägigen wissenschaftlichen Arbeiten entnommen (Masood u.a. 2022; Mirsky/Lee 2022; Nguyen u.a. 2022; Pawelec/Bieß 2021):

artificial intelligence-synthesized content | deep fake | deep fake% | deepfake | deep-fake | deep-fake% | deepfake% | Deepfake-Video% | face manipulation% | face spoof% | face swap% | face synthesis% | face-swap% | face clon% | facial reenactment% | facial synthesis% | Gesicht-Morphing% | Morphing face% | speech synthe% | Sprache% klonen% | Sprache% synthetisieren% | Sprachsynthetisierung % | Stimme% klonen% | Stimme% synthetisieren% | synthesized facial image% | synthetic media% | voice clon% | voice manipulation% | Voice spoof% | voice swap%

[7] Das Prozentzeichen stellt einen Platzhalter (sog. Wildcard) dar, mit dem auch Wörter gefunden werden, die weitere Buchstaben nach dem eigentlich gesuchten Wort einschliessen. Beispiel: «deepfake%» würde auch den Plural deepfakes einschliessen oder «face swap%» auch face swapping finden. Der vertikale Strich (eng. Pipe) bedeutet ein logisches «Oder».

### 2.8.2. Resultate

Erster Suchlauf (drei Schlagworte):

Insgesamt wurden über alle Jahre und Länder hinweg 562 relevante Publikationen gefunden. Wie in Abbildung 9 zu sehen, wurde ein Grossteil dieser Publikationen von Autorinnen und Autoren aus den USA, China und Indien veröffentlicht. Die meisten Publikationen (27 %) entfallen auf die USA, dicht gefolgt von China (26 %) und Indien mit 14 %. Auf Grossbritannien entfallen noch knapp 6 %. Das erste Land aus der EU ist Italien mit 4 %. Aus der Schweiz kommen etwas mehr als 1 % der Publikationen. Alle EU-Länder inklusive der Schweiz und Grossbritannien (EU+CH+GB) kommen auf insgesamt 86 Publikationen (22 %).

Abbildung 9 zeigt zudem die Anzahl der Publikationen zum Thema Deepfakes im Vergleich zu der Gesamtanzahl aller Publikationen des jeweiligen Landes. Hier zeigt sich ein weitgehend homogenes Bild (0,001 bis 0,0015 %) für die meisten Länder. Auffallend ist nur Indien, in welchem doppelt so viele (0,002 %) aller Publikationen zum Thema Deepfake gehören.

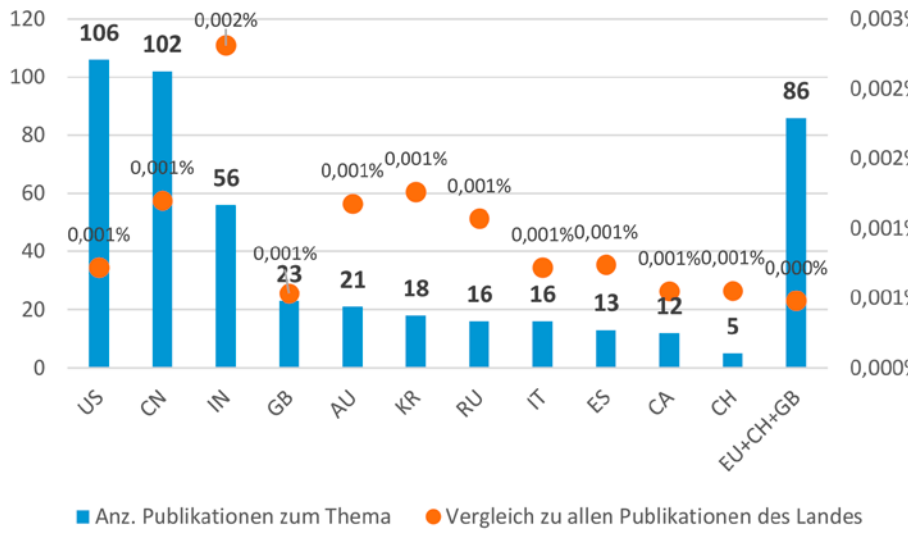


**Abbildung 9:** Top-10-Anzahl der Veröffentlichungen im ersten Suchlauf pro Land sowie die Anzahl der Publikationen im Vergleich zur gesamten Publikationstätigkeit des Landes (EU+CH+GB umfasst alle EU-Staaten, Grossbritannien und die Schweiz)

Ein Vergleich der Publikationstätigkeiten gruppiert nach Ländern (Top-5-Länder und die Schweiz) im zeitlichen Verlauf wird in Abbildung 10 dargestellt.[8] Hierbei wird deutlich, dass der Begriff Deepfake in Publikationen erst 2019 auftauchte und die Anzahl der Veröffentlichungen danach stark anstieg. Schweizer Wissenschaftler veröffentlichten hingegen im Jahr 2021 erste Arbeiten mit diesem Schlagwort.

Eine Analyse der zeitlichen Verteilung aller Publikationen weltweit zeigt ein ähnliches Bild. Im Jahr 2019 gab es 24 Veröffentlichungen, im Jahr darauf stieg dieser Wert um 425 % auf 126 Werke. Der Höhepunkt liegt bisher im Jahr 2021 mit 243 Veröffentlichungen.

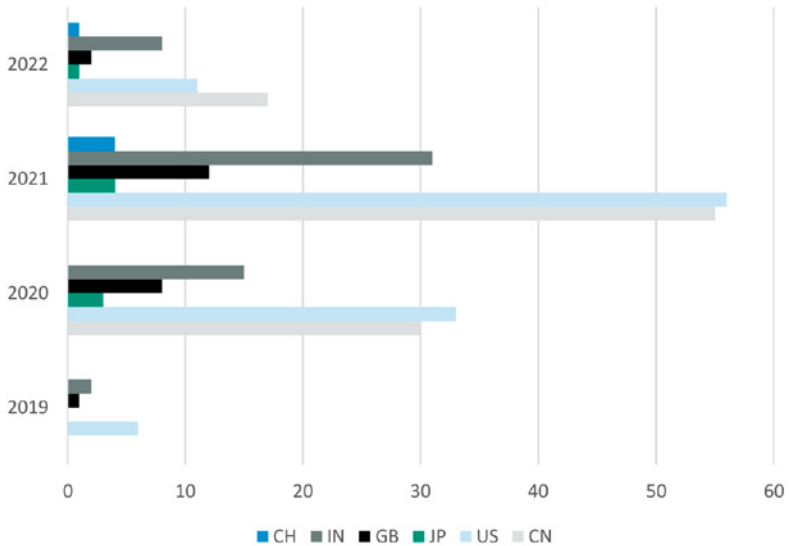


**Abbildung 10:** Anzahl der Publikationen gruppiert nach Land und Jahr für den ersten Suchlauf

Erweiterter Suchlauf (mit allen Schlagworten):

Insgesamt wurden über alle Jahre und Länder hinweg 3658 relevante Publikationen gefunden. Wie in Abbildung 11 zu sehen, wurde ein Grossteil dieser

[8] Der enorme Rückgang im Jahr 2022 ist auf die Datenbasis zurückzuführen, welche nur Publikationen bis zur 17. Kalenderwoche des Jahres 2022 beinhaltet.

Publikationen von Autorinnen und Autoren aus China, den USA und Japan veröffentlicht. Die meisten Publikationen (knapp 16 %) entfallen auf China, gefolgt von den USA und Japan jeweils mit ca. 12 %. Auf Grossbritannien entfallen 8 % und auf Indien 6 %. Aus Deutschland stammen 3,4 %, aus Frankreich etwas weniger als 3 % und aus der Schweiz ca. 1 % aller Publikationen. Alle EU-Länder, die Schweiz und Grossbritannien (EU+CH+GB) kommen auf insgesamt 880 Publikationen (ca. 24 %).

Abbildung 11 zeigt zudem die Anzahl der Publikationen zum Thema Deepfakes im Vergleich zu der Gesamtanzahl aller Publikationen des jeweiligen Landes. Hier zeigt sich ein weitgehend homogenes Bild, jedoch mit zwei Ausreissern. Während bei den meisten Ländern der prozentuale Anteil von Publikationen zum Thema Deepfakes im Vergleich zu allen Publikationen des Landes zwischen 0,003 und 0,009 % schwankt, zeigt sich in Japan (0,013 %) und Tschechien (0,019 %) ein höheres relatives Publikationsaufkommen.

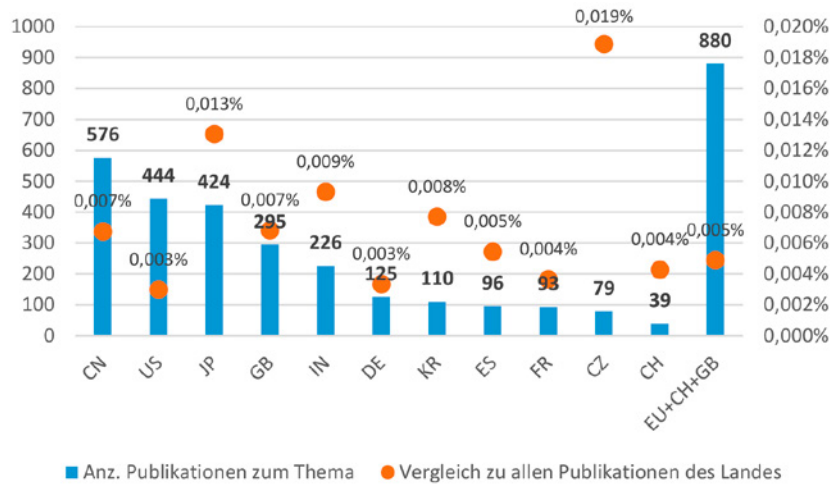


**Abbildung 11:** Top-10-Anzahl der Veröffentlichungen im erweiterten Suchlauf pro Land sowie die Anzahl der Publikationen im Vergleich zur gesamten Publikationstätigkeit des Landes (EU+CH+GB umfasst alle EU-Staaten und die Schweiz)

Ein Vergleich der Publikationstätigkeiten gruppiert nach Ländern (Top-5-Länder und die Schweiz) im zeitlichen Verlauf wird in Abbildung 12 dargestellt.[9] Im

[9] Der enorme Rückgang im Jahr 2022 ist auf die Datenbasis zurückzuführen, welche nur Publikationen bis zur 17. Kalenderwoche des Jahres 2022 beinhaltet.

Gegensatz zum ersten Suchlauf sind die Ergebnisse hier vielfältiger und Publikationen zu thematisch verwandten Schlagworten sind dominant. Dennoch wird deutlich, dass ein Grossteil der Publikationen erst in den letzten vier Jahren entstand. Ab 2018 nahmen die Publikationen stetig zu, während sie in den Jahren zuvor eher gleichmässig verteilt waren. Die Publikationen aus der Zeit vor 2018 beschäftigten sich zum Grossteil mit dem Schlagwort «Speech synthesis» (2328 Publikationen). In der Schweiz erschienen die ersten Publikationen zum Thema 2009, während in Japan und China augenscheinlich bereits 2004 das Thema erforscht wurde. Eine chinesische Dominanz auf dem Gebiet der Publikationen besteht indes schon immer.

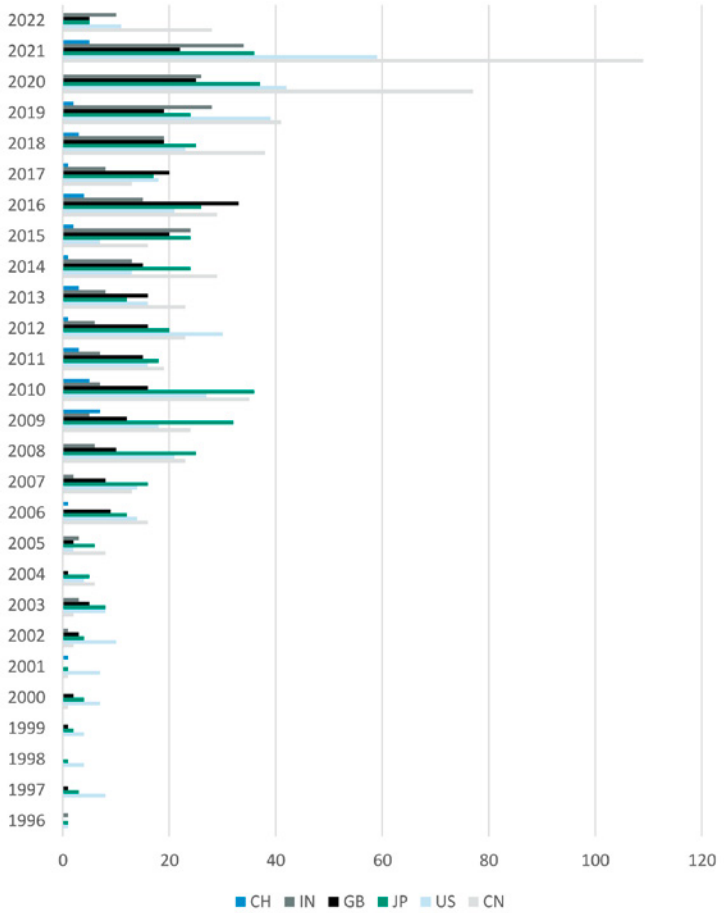


**Abbildung 12:** Anzahl der Publikationen gruppiert nach Land und Jahr für den erweiterten Suchlauf

Eine Analyse der zeitlichen Verteilung aller Publikationen weltweit von 1996 bis 2022 zeigt ein ähnliches Bild. So gab es 1996 nur 4 Publikationen, während dieser Wert innerhalb von einem Jahr um 325 % (17 Publikationen) anstieg. In den Folgejahren schwankten die Werte weniger stark. Im Jahr 2007 wurde dann zum ersten Mal die Hundertermarke überschritten (103 Publikationen). Das Maximum wurde 2021 mit 370 Publikationen erreicht.

#### 2.8.2.1. Qualitative Auswertung der Ergebnisse

In diesem Kapitel wird die quantitative Auswertung der beiden obigen Kapitel in einen qualitativen Zusammenhang gebracht. Die relevanten Fragestellungen hierzu sind:

1. Wie ist die Forschung international verknüpft?
2. Auf welche Themen innerhalb der Deepfake-Forschung wird sich in welchem Land fokussiert?
3. In welchen Wissenschaftsgebieten wird an dem Thema gearbeitet und sind in den Veröffentlichungen unterschiedliche Gebiete einbezogen?

Zur Beantwortung der Fragen wurden Publikationsinformationen wie Titel, Abstrakt und Schlüsselworte sowie Autoreninformationen ausgewertet.

Die Gruppierung und Zählung von Autoren-Herkunftsland pro Publikationen zeigen, dass fast alle Paper von Autorinnen und Autoren aus demselben Land stammen. Eine nennenswerte internationale Zusammenarbeit in der Forschung ist daher nicht zu erkennen. Nur sehr vereinzelt sind gemeinsame Publikationen von Autoren unterschiedlicher Länder anzutreffen (z.B. China und Singapur, Belgien und Schweden oder Grossbritannien und den Niederlanden). Dabei fällt auf, dass es sich dabei um Staaten handelt, die eine kulturelle Nähe zueinander aufweisen. Im Onlineappendix[10] finden Sie die ganze Tabelle mit der aufsummierten Anzahl von Ländern pro Veröffentlichung.

Bei der Untersuchung von unterschiedlichen Themenfeldern in den einzelnen Ländern war von Interesse, in welchen Ländern mehr über die Erstellung geforscht wird und in welchen es eher um die Erkennung geht.[11] Die Analyse der

[10] https://zenodo.org/records/10847935.

[11] Anmerkung: Die Daten basieren auf dem Land des Autors. Es könnte jedoch auch möglich sein, dass manche Autoren nicht mehr in ihrem Heimatland ansässig sind. Dann würde aber trotzdem in die Auswertung das Herkunftsland des Autors einbezogen.

Schlagworte zeigt, dass basierend auf unseren Schlagworten,[12] grundsätzlich mehr über die Erkennung als über die Erstellung publiziert wird. Der Blick auf unterschiedliche Länder (Tabelle 3: Anzahl der Veröffentlichungen zum Thema Erkennung oder Erstellung pro Land) zeigt, dass vor allem die Schweiz und andere europäische Länder sehr viel mehr zur Erkennung forschen. Am geringsten fällt dieser Unterschied in Indien auf, wo nur 39 % mehr zum Thema Erkennung geforscht wird.

Die Frage, in welchen unterschiedlichen Wissenschaftsgebieten an der Deepfake-Forschung gearbeitet wird, lässt sich über die angegebenen Forschungsgebiete der Autorinnen und Autoren herausfinden. Hier zeigt sich ein eher heterogenes Bild (eine komplette Übersicht finden Sie im Onlineappendix[13]). Ein Grossteil der Arbeiten ist von Autorinnen und Autoren mit vier verschiedenen Forschungsfeldern geschrieben (23 %), gefolgt von drei Feldern (21 %) und zwei Feldern (19 %). Nur in einem Forschungsfeld sind 14 % der Veröffentlichungen. Auf der anderen Seite gibt es aber auch vier Publikationen (9 %), die zehn oder mehr Forschungsfelder aufweist. Konkret wird am meisten im Feld «Computer Science» dazu geforscht (86 % der Autoren aller Paper sind in diesem Bereich tätig), gefolgt von «Engineering» (39 %) und «Social Sciences» (22 %).

**Tabelle 3:** Anzahl der Veröffentlichungen zum Thema Erkennung oder Erstellung pro Land

| Ausgewähltes Land | Anzahl Nennungen für | | Unterschied zwischen ES und EK | |
|---|---|---|---|---|
| | Erstellung (ES) | Erkennung (EK) | | |
| US | 18 | 42 | + 24 | 133 % |
| CN | 26 | 70 | + 44 | 169 % |
| IN | 23 | 32 | + 9 | 39 % |
| CH | 0 | 2 | + 2 | 200 % |
| DE | 0 | 3 | + 3 | 300 % |
| FR | 1 | 5 | + 4 | 400 % |

[12] Anmerkung: Die Ergebnisse basieren auf den gesuchten Schlagworten. Bei einer anderen Auswahl der Schlagworte könnten sich die Ergebnisse verändern.
Erkennung: Die Verwendung der Schlagworte «detection|detect|forensic|recognition|analysis|attention» führt zu 245 Publikationen.
Erstellung: Die Verwendung der Schlagworte «generator|generative|creation|create|creating|manipulation|forgery|tampering|tamper» führt zu 124 Publikationen.

[13] https://zenodo.org/records/10847958.

## 2.9. Ist- und Trendanalyse: Zwischenfazit

Im Rahmen der Ist- und Trendanalyse wurde der Stand der Forschung zu Deepfake-Technologien aufgearbeitet. Erstens wurde erläutert, welche technischen Möglichkeiten zur Erschaffung von Deepfakes gegenwärtig existieren und in naher Zukunft absehbar sind. Zweitens wird dargestellt, welche technischen Möglichkeiten zur Unterscheidung von echten und synthetischen bzw. manipulierten Inhalten bestehen.

Die Untersuchung der aktuellen und absehbaren technischen Möglichkeiten zur Erschaffung von Deepfakes zeigte, dass in den vergangenen Jahren beachtliche Erfolge in der KI-basierten Synthetisierung und Manipulation von Bild-, Audio- und Textinhalten erzielt wurden.

Im Bereich bildbasierter Deepfakes stehen gegenwärtig fünf Techniken zur Verfügung:

- *Facial reenactment*, die Manipulation des Gesichtsausdrucks
- *Face morphing*, die Verschmelzung von mehreren Gesichtern
- *Face swapping*, das Tauschen oder Ersetzen eines Gesichts mit einem anderen Gesicht
- *Gesichtsgeneration*, die Erschaffung von Gesichtern, die so in der Realität nicht existieren, und
- *Full body puppetry*, das sog. Ganzkörperpuppenspiel, bei dem die Pose oder Bewegungen eines Körperteils oder des gesamten Körpers verändert werden.

Im Bereich von Deepfake-Audios hat sich die Entwicklung in den vergangenen Jahren einerseits in Richtung zunehmend glaubwürdiger Resultate entwickelt, die zudem andererseits mit zunehmend weniger Trainingsmaterial erzielt werden können. Eine vergleichbare Entwicklung hat auch im Bereich von Text-Generatoren stattgefunden. Neben der mittlerweile guten Qualität des Text-Outputs bieten moderne Textgeneratoren inzwischen auch das Kopieren des Schreibstils einer spezifischen Person.

Einerseits ist bereits heute die Erschaffung einer breiten Palette an Deepfake-Inhalten möglich. Ausgehend von der Imitation der Schreibweise und Sprache, erlauben moderne Deepfake-Technologien auch das Kopieren der Sprechweise, des Gesichtsausdrucks und sogar der Körperbewegungen eines Menschen –

immer vorausgesetzt, die Erschaffer verfügen über ausreichend Inputdaten und Computerleistung. Zudem lassen sich in zunehmend besserer Qualität Hintergrund bzw. Szenerie vollständig synthetisieren und verändern. Damit ist nicht nur die Erschaffung von Deepfakes möglich, die zwingenderweise Menschen abbilden, sondern auch jeder erdenklichen sonstigen Situation, etwa ein gefälschter Flugzeugabsturz oder eine gefälschte Naturkatastrophe.

Andererseits bestätigen von uns selbst durchgeführte Versuche zur Erstellung von Deepfakes eher eine vorsichtig positive Bewertung.[14] Anwendungen zur Deepfake-Erschaffung werden zwar qualitativ zunehmend hochwertiger und auch einfacher in der Bedienung, erfordern allerdings weiterhin grosses technisches Know-how und Rechenkapazitäten bzw. Zeit und Geld.

Ändern könnte sich diese Situation, wenn entweder Video-Deepfake-Generatoren einfacher nutzbar werden oder sobald Text-zu-Video-Verfahren ausgereift und für breite Massen der Bevölkerung verfügbar sind. Mit solchen Generatoren wäre es möglich, beliebige Deepfake-Videos unter Eingabe von Textbefehlen zu erstellen – so wie es etwa mit ChatGPT im Bereich von KI-generierten Texten oder mit KI-basierten Bildgeneratoren wie Midjourney und Dall-E bereits heute möglich ist. Ähnlich wie bei aktuellen Text- und Bildgeneratoren kann davon ausgegangen werden, dass die bekanntesten Programme ihrer Art Schutzvorkehrungen treffen werden, um schädliche und unethische Nutzungen auszuschliessen. Andererseits können solche Schutzvorkehrungen umgangen werden, und die Entstehung weiterer Generatoren, die ohne ethische Leitlinien programmiert werden, ist naheliegend. Auch wenn also keine akute Gefahr der massenweisen Erstellung und Flutung der sozialen Medien mit Deepfake-Videos droht, wird die Erschaffung von glaubwürdigen Deepfake-Videos in der Zukunft deutlich einfacher werden. Wie die Diskussionen in den folgenden Kapiteln zeigen werden, braucht es für einen missbräuchlichen Effekt keine Masse an Deepfakes. Auch ein einzelnes Deepfake kann grossen Schaden anrichten. Und wie die Reaktionen auf den Papst in Daunenjacke demonstriert haben, ist nicht einmal eine Irreführungs- oder Schädigungsabsicht notwendig: Menschen können auch von Medieninhalten in die Irre geführt werden, die zu Unterhaltungszwecken erstellt wurden.

Technische Massnahmen, die als Abhilfe diskutiert werden, sind Methoden zur Authentifizierung von originalen Inhalten, zur Kennzeichnung von Deepfake-Inhalten sowie Detektionstools zur Erkennung von Deepfake-Videos. Keine die-

[14] https://zenodo.org/records/10847968.

ser Massnahmen ist ohne Schwächen. Gegen die Authentifizierung vertrauenswürdiger Inhalte und die Kennzeichnung von Deepfake-Inhalten spricht erstens ihre prinzipielle technische Überwindbarkeit. Ein fälschlich als vertrauenswürdig signiertes Deepfake könnten nicht nur selbst grossen Schaden anrichten, sondern auch die Vertrauenswürdigkeit signierter Dokumente grundsätzlich beschädigen. Die Überwindung solcher Schutzvorkehrungen könnte vermutlich nur besonders ressourcenstarken (staatlichen) Akteuren gelingen. Dies könnte im Umkehrschluss aber auch bedeuten, dass die Masse der kursierenden Deepfakes gekennzeichnet und die Masse der vertrauenswürdigen Inhalte als solches signiert wäre. Zweitens können Authentifizierungsmassnahmen wie digitale Signaturen mit einer Form der Totalüberwachung einhergehen, sodass die Anonymität der erstellenden Personen nicht mehr gewährleistet wäre. Dies könnte z.B. Menschenrechtsaktivisten in diktatorischen oder autoritären Regimen oder auch Whistleblower gefährden.

Angesichts der Investitionen in die Entwicklung von Detektoren und den technischen Herausforderungen, denen GANs (aber auch andere Deepfake-Technologien) gegenüberstehen, lässt sich derzeit zwar nicht sagen, wie gut Detektoren in der Zukunft bei der Erkennung von Deepfake-Inhalten sein werden. Unsere Tests mit Detektoren als auch der Stand der Literatur zeigen jedoch, dass frei verfügbare Deepfake-Detektoren zumindest derzeit nicht zuverlässig sind.

Im letzten Schritt wurde eine bibliometrische Auswertung wissenschaftlicher Publikationen durchgeführt, um einen Überblick über das Publikationsaufkommen und die Positionierung der Schweiz im internationalen Vergleich zu erhalten. Die Ergebnisse verdeutlichen die führende Rolle der Volksrepublik China und der Vereinigten Staaten im Bereich der technisch-angewandten Forschung zu Deepfakes. Unter Einbeziehung auch der Grundlagenforschung zu Deepfake-Technologien reduzierte sich der zwischenstaatliche Abstand hingegen deutlich. Das Publikationsniveau in der Schweiz bewegt sich sowohl im Hinblick auf die absolute Zahl als auch auf den prozentualen Anteil von Deepfakes betreffenden Publikationen zur Gesamtpublikationstätigkeit im Durchschnitt der europäischen Staaten. Die Analyse zeigte auch, dass der Grossteil der Forschenden zu Deepfakes aus dem Bereich der Informatik stammen und dass die Erkennung von Deepfakes stärker erforscht wird als deren Produktion.

# 3. Wahrnehmung von Deepfakes in der Schweizer Bevölkerung

*Daniel Vogler, Adrian Rauchfleisch & Gabriele de Seta*

## 3.1. Theorie und Forschungsstand

Sogenannte Deepfakes oder synthetische Medien sind ein relativ neues Phänomen des Internets. Der Begriff synthetische Medien bezeichnet Bilder, Ton oder Videos, die durch KI-Technologien generiert wurden (Westerlund 2019). Diese Medien werden als vielversprechend für Bildungs- und Unterhaltungszwecke angesehen. Sie stellen aber auch eine potenzielle Bedrohung für demokratische Gesellschaften dar (Ahmed 2021b; Hameleers u.a. 2022; Vaccari/Chadwick 2020; Fallis 2021), da sie eine Manipulation von echten Medieninhalten beinhalten oder von Grund auf komplett gefälscht sein können, mit dem Ziel, die Zuschauerinnen und Zuschauer zu täuschen. Dies kann beispielsweise ein gefälschtes Video eines Politikers oder einer Berühmtheit sein, das etwas zeigt, was er oder sie nie getan oder gesagt hat (vgl. zu Deepfake-Szenarien Kapitel 6 und 7). Daher wird die Fähigkeit, synthetische und manipulierte Medien zu erkennen und von echten Medien zu unterscheiden, zu einer wichtigen Fähigkeit, welche die allgemeine Internet- und Medienkompetenz der Menschen ergänzen soll (Hwang u.a. 2021; McCosker 2022; Shin/Lee 2022). Zurzeit gibt es jedoch nur wenig Forschung darüber, welche Erfahrungen Menschen in der Schweiz mit Deepfakes haben, wie sie Chancen und Risiken von Deepfake-Technologien wahrnehmen und ob sie in der Lage sind, synthetische Medien zu erkennen, und welche Rolle digitale Medienkompetenz in diesem Prozess spielt. Die drei zentralen Forschungsfragen (FF) dieser Studie lauten deshalb:

– **FF 2.1:** Welche Erfahrungen haben Menschen in der Schweiz bislang mit Deepfakes?

– **FF 2.2:** Werden Deepfakes in der Schweizer Bevölkerung eher als Chance oder als Risiko wahrgenommen und welche Faktoren beeinflussen diese Wahrnehmung?

– **FF 2.3:** Sind Menschen in der Schweiz in der Lage, Deepfakes zu erkennen, und welche Faktoren beeinflussen diesen Prozess?

Mit der Studie werden umfangreiche deskriptive Daten zu Deepfakes in der Schweiz erhoben. Aus den Resultaten werden zusätzlich erste Ideen für Massnahmen im Umgang mit Deepfakes abgeleitet, etwa in der Ausbildung.

### 3.1.1. Erfahrung mit Deepfakes

Da Deepfakes ein relativ neues Phänomen sind, ist die Forschung noch lückenhaft. Insbesondere Studien zur öffentlichen Wahrnehmung sind kaum vorhanden. Zur Schweiz gibt es noch keine gesamtheitliche Studie zur Wahrnehmung von Deepfakes in der Bevölkerung. Wir erheben deshalb in einem ersten Schritt deskriptive Daten zur Schweiz. Wir untersuchen die Erfahrungen von Schweizerinnen und Schweizern mit Deepfakes und welche Unterschiede zwischen soziodemografischen Gruppen bestehen. Konkret erheben wir, ob Menschen in der Schweiz schon von Deepfakes gehört haben, ob sie schon Deepfakes gesehen haben und ob sie schon Deepfakes erstellt und weiterverbreitet haben. Forschungsergebnisse zeigen einen starken Einfluss soziodemografischer Variablen auf den Umgang mit (Medien-)Technologien im Allgemeinen (Gustafsod 1998; Störk-Biber u.a. 2020) und Deepfakes im Besonderen (Cochran/Napshin 2021; Wang/Kim 2022). Daher untersuchen wir den Einfluss von Geschlecht, Alter und Bildung auf die Wahrnehmung von Deepfakes.

Im Prozess der Wahrnehmung von Technologie spielt Kommunikation off- wie online eine tragende Rolle. Dazu gehören Gespräche mit Freunden oder Familie, der Austausch in grösseren Gruppen über Messengerdienste oder öffentliche Debatten auf Social-Media-Plattformen. Viele Menschen kommen mit Deepfakes auf Social Media oder Messengerdiensten in Kontakt, wenn sie beispielsweise Videos von Peers erhalten und diese dann auch selber weiterverbreiten (Hameleers u.a. 2022; Vaccari/Chadwick 2020). Eine besonders wichtige Rolle spielen aber nach wie vor journalistische Medien (Gosse/Burkell 2020). Mit der Themenselektion, der Fokussierung auf bestimmte Aspekte des Themas, Akteuren, die in den Beiträgen zu Worte kommen, und dem verwendeten Framing können Medien die Wahrnehmung von Technologie mitbestimmen. Wenn also in der Berichterstattung zu Deepfakes der Aspekt der Desinformation im Zentrum steht, also ein negatives Framing dominiert, sind zwangsläufig auch negative Wahrnehmungseffekte anzunehmen. Im Fall von Deepfakes ist der Link zum Thema Desinformation oder Fake News in der Berichterstattung (Gosse/Burkell 2020; Wahl-Jorgensen/Carlson 2021; Yadlin-Segal/Oppenheim 2021) und im Schweizer Journalismus stark ausgeprägt (siehe Kapitel 2). Wir untersuchen deshalb, auf welchen Kanälen Menschen mit Deepfakes in Berührung kommen und ob

ein Zusammenhang zwischen der Mediennutzung und dem Kontakt mit Deepfake-Technologie besteht.

### 3.1.2. Chancen- und Risikowahrnehmung von Deepfakes

Wie Menschen Technologie wahrnehmen und damit umgehen, ist eine etablierte Forschungsfrage, insbesondere mit Fokus auf deren Risiken und Chancen (Binder u.a. 2012; Covello 1983). Dasselbe gilt für neue Medien- und Kommunikationstechnologien wie zum Beispiel Deepfakes (Cochran/Napshin 2021; Kleine 2022; Lu/Chu 2023) oder mobile Apps (Tay u.a. 2021). Von der Wahrnehmung in der Gesellschaft hängt es massgeblich ab, ob und wie sich neue Technologien in einer Gesellschaft etablieren können. Beispielsweise sind technologische Fortschritte im Bereich der Gentechnologie (Frewer u.a. 1999), Nuklearenergie (Siegrist/Visschers 2013) oder Impfungen (Wilson u.a. 2015) bei Teilen der Bevölkerung auf Skepsis gestossen.

Die Wahrnehmung beeinflusst die Akzeptanz und im weiteren Verlauf auch die Art und Weise, wie Menschen Technologien im Alltag nutzen, primär als Konsumenten, aber auch als Anwender (Slovic 1999). Die Wahrnehmung ist daher eine zentrale Komponente, die den Umgang einer Gesellschaft mit neuen Technologien wie künstlicher Intelligenz oder Deepfakes beeinflusst (Bareis/Katzenbach 2022). In der Regel wird zwischen Risiken und Chancen von neuen Technologien unterschieden (Siegrist/Visschers 2013). In Bezug auf Deepfakes interessiert uns, ob Risiken die Wahrnehmung dominieren oder eine positive Wahrnehmung, die eher Chancen der Technologie sieht, dominiert. Diese Wahrnehmung hängt oftmals von individuellen Faktoren ab. Dazu gehören typischerweise Alter, Geschlecht und Bildung (Störk-Biber u.a. 2020), die wir auch in unserer Studie untersuchen.

Die Wahrnehmung von Technologie hängt auch von den direkten Erfahrungen oder Einstellungen der Menschen mit dieser Technologie ab. Wir untersuchen deshalb, ob Menschen, die bereits Erfahrung mit Deepfake-Technologien haben, also schon Deepfakes gesehen, selber erstellt oder weiterverbreitet haben, eine geringere Risikowahrnehmung aufweisen als Personen, die nur eine vage Vorstellung der Technologie haben und Deepfakes lediglich durch einen kritischen Mediendiskurs kennen. Wir untersuchen zudem eine Reihe weiterer individueller Einstellungen gegenüber Deepfakes. Erstens die selbsteingeschätzte Deepfake-Erkennungskompetenz, also wie gut man nach eigener Einschätzung in der Lage ist, Deepfakes zu erkennen. Zweitens, wie stark die Befragten den

Effekt von Deepfakes auf ihre eigene Meinung und die Meinung anderer einschätzen. Drittens, die wahrgenommene Verbreitung von Deepfakes, also ob Deepfakes nach der Einschätzung der Befragten viel oder wenig verbreitet sind. Viertens, das individuelle Vertrauen in Politik, Medien und Wirtschaft.

Für die Risikowahrnehmung spielt der Anwendungsbereich von Deepfakes eine zentrale Rolle. Die Verwendung von Deepfakes zur Unterhaltung, beispielsweise mit Tools wie Midjourney oder FaceSwap, wird vermutlich stärker mit Chancen und individuellen Gratifikationen als Risiken in Verbindung gebracht. Im Bereich von gesellschaftspolitisch relevanten Informationen, beispielsweise zu Politik oder Gesundheit, dürfte hingegen deutlich die Wahrnehmung als Risiko im Zentrum stehen (Hameleers u.a. 2022; Vaccari/Chadwick 2020). Besonders prominent wird Desinformation in der politischen Kommunikation während Wahlen oder Abstimmungen oder im Gesundheitsbereich (z.B. zu Impfungen) behandelt. Wir gehen davon aus, dass die Risikowahrnehmung bei den Befragten davon abhängt, welcher Gesellschaftsbereich angesprochen wird. Deshalb untersuchen wir die Wahrnehmung von Chancen und Risiken von Deepfake-Technologien separat für Politik, Medien, Wirtschaft und das Individuum.

Studien zeigen, dass das Framing von Technologien deren Wahrnehmung beeinflusst (Tay u.a. 2021). Weiter gehen wir deshalb davon aus, dass es eine Rolle spielt, wie eine Technologie bezeichnet wird. Die Bezeichnung Deepfake ist per se mit einer negativen Bewertung verbunden, weil es einerseits den negativ konnotierten Term «Fake» enthält und andererseits in der öffentlichen Debatte bereits negativ aufgeladen wurde, während die Bezeichnung synthetische Medien als neutrale Bezeichnung mit einer geringeren Risikowahrnehmung verbunden ist. Uns interessiert deshalb, ob die Verwendung der Bezeichnung Deepfake die Risikokomponente der Wahrnehmung in der Bevölkerung verstärkt und mögliche positive Aspekte dadurch verdrängt.

### 3.1.3. Erkennen von Deepfakes

Um die negativen Aspekte von Deepfakes zu bekämpfen, werden in der Literatur drei Massnahmen diskutiert: staatliche Regulierung, technologische Massnahmen zur Erkennung und (digitale) Medienkompetenz (McCosker 2022). Grossflächige Regulierungen und Verbote von Deepfake-Technologien sind rechtlich kaum umsetzbar und auch dysfunktional, beispielsweise aufgrund von wirtschaftlichen Standortnachteilen (Bareis/Katzenbach 2022). Technologische Massnahmen zur Erkennung von Deepfakes können helfen, allerdings sind sie

noch wenig ausgereift und hinken – analog zu Fact-Checking von Desinformation – der neuesten Entwicklung in der Produktion hinterher (vgl. Kapitel 2.6 und 2.7). Deshalb kommt der Medienkompetenz eine besondere Bedeutung zu. Obwohl auch diese Massnahme in der Umsetzung anspruchsvoll ist (Wagner/Blewer 2019), kann sie die Bürgerinnen und Bürger befähigen, mit den Risiken und den Potenzialen von synthetischen Medien besser umzugehen. Diese Massnahmen stärken gleichzeitig die Resilienz einer Gesellschaft gegen negative Effekte von Deepfakes sowie die Kompetenz der verantwortungsvollen Nutzung der Technologie. Risiken sind in diesem Zusammenhang von zentraler Bedeutung, da Bildmaterial nach wie vor als besonders glaubhaft wahrgenommen wird. Dieser Effekt wird gerade im Zusammenhang mit Desinformation als besonders problematisch angesehen (Lee/Shin 2022).

Aktuell ist man in der Lage, qualitativ hochwertige synthetische Medien herzustellen, die kaum von echten Bildern unterscheidbar sind. Mit Blick auf die Zukunft wird es für Menschen immer schwieriger werden, zwischen synthetisch erzeugten Medien und echtem Bildmaterial zu unterscheiden (Godulla u.a. 2021; Fallis 2021). Daher sind Strategien jenseits von optischem Erkennen von Deepfakes notwendig. Ähnlich wie bei der Erkennung von Desinformation wird allgemeiner Medienkompetenz zur Erkennung von Deepfakes eine positive Rolle zugesprochen (Hwang et al., 2021). Dies umfasst einerseits einen kritischen Umgang mit Information sowie das Prüfen von Quellen und Absendern. Andererseits kann auch Wissen über Technologien, etwa ihrer Potenziale und Schwachstellen, sowie Produktionsprozesse hilfreich sein. Hier sind spezifische Fähigkeiten notwendig, zum Beispiel Kenntnisse über Algorithmen und künstliche Intelligenz, also ähnliche Fähigkeiten, die im Umgang mit dem Internet oder Social Media erlernt wurden, sog. Internet-Skills (Hargittai u.a. 2019; Hargittai/Hsieh 2012) oder New Media Skills (Koc/Barut 2016; Tandoc Jr u.a. 2021). Dazu gehören neben Kenntnissen von technischen Merkmalen und Prozessen auch soziokulturelle Aspekte von neuen Medien (Koc/Barut 2016). In der Literatur finden sich auch Forderungen und erste Ansätze zur Deepfake-Literacy (Ali u.a. 2021; Lee/Shin 2022; McCosker 2022; Wagner/Blewer 2019).

*Literacy-Intervention*

Wir gehen davon aus, dass Menschen, die eine kurze Anleitung zur Erkennung von Deepfakes erhalten, besser in der Lage sind, Deepfakes zu erkennen (Hwang u.a. 2021). Normalerweise sind solche Literacy Interventions aufwendig und eher langfristig ausgerichtet (Wagner/Blewer 2019). Dennoch zeigen

Studien aus dem Bereich der Desinformationsforschung, dass kurze Literacy-Interventionen einen substanziellen Einfluss auf die Fähigkeit zur Erkennung von Desinformation haben können (Guess u.a. 2020). Daher testen wir in einem Survey-Experiment, ob eine Literacy-Intervention (Potter 2013) einen positiven Effekt auf die Fähigkeit hat, Deepfakes zu erkennen.

*Medienkompetenz*

Zudem erwarten wir, dass neben der situativen Literacy-Intervention erlernte allgemeine Medien- und Internetkompetenzen (Hargittai/Hsieh 2012; Hargittai u.a. 2019; Koc/Barut 2016) einen positiven Effekt auf das Erkennen von Deepfakes haben. Durch die Thematisierung und Einordnung von Deepfakes und der zugrunde liegenden Technologien können journalistische Medien dazu beitragen, Wissen zu Deepfakes zu vermitteln (siehe Kapitel 2). Forschungsergebnisse zeigen beispielsweise, dass die Nutzung von journalistischen Medien die Resilienz von Menschen gegen Desinformation stärkt (Humprecht u.a. 2021). Social Media (z.B. TikTok, YouTube, Facebook) und Messenger (z.B. WhatsApp oder Telegram) nehmen in der Verbreitung von Deepfakes eine zentrale Rolle ein. Auf diesen Kanälen treffen Menschen am ehesten auf Deepfakes (Hameleers u.a. 2022) und können Erfahrungen und Wissen im Umgang mit der Technologie sammeln. Wir untersuchen deshalb, ob allgemeine Medienkompetenz, die *Social-Media-Literacy* sowie *Internet-Skills* einen Einfluss auf die Deepfake-Erkennungskompetenz haben.

*Exposure und individuelle Erfahrungen*

Die Forschung zeigt, dass das Wissen über Deepfakes und die Erfahrungen, die Personen bereits mit Deepfakes gemacht haben, sich positiv auf die Erkennung von Deepfakes auswirken (Shin/Lee 2022). Wir gehen davon aus, dass Menschen, die bereits Erfahrungen mit Deepfakes gemacht haben und sich damit Wissen über die Technologien angeeignet haben, besser in der Lage sind, Deepfakes zu erkennen. Die wenigen vorhandenen Studien legen zudem nahe, dass soziodemografische Variablen eine wichtige Rolle bei der Erkennung von Deepfakes spielen. Wir untersuchen daher Einfluss von Geschlecht, Alter und Bildung auf die Erkennungskompetenz.

## 3.2. Methodische Vorgehensweise

Um unsere Forschungsfragen zu beantworten, wurden zwei Befragungen konzipiert, in denen neben einer generellen Befragung jeweils ein Onlineexperiment eingebaut wurde. Die Rekrutierung der Teilnehmenden aus der Deutschschweiz und der Suisse Romande erfolgte über das Panel eines externen Anbieters (Bilendi). Die Befragungen wurden beide mit der Software Unipark programmiert. Das Sample ist jeweils in den Grundzügen (Alter, Geschlecht, Sprachregion und Bildung) repräsentativ für die Schweizer Bevölkerung (Deutschschweiz und Suisse Romande), beinhaltet aber nur Personen, die das Internet nutzen.

### 3.2.1. Vorstudie und Pretest

Die Vorstudie (n = 660) haben wir im Juni 2023 durchgeführt. Sie diente insbesondere Dazu, das verwendete Videomaterial zu testen. Dies umfasste das Testen der technischen Aspekte sowie das Ermitteln des Schwierigkeitsgrads der Videos. Neben dem Pretest und dem Test der Stimuli (Beispiele von Deepfakes) wurde ein erstes Experiment durchgeführt. In diesem Experiment variierten wir die Bezeichnung unseres zentralen Untersuchungsgegenstandes zwischen «Deepfake» und «synthetische Medien». Eine Gruppe erhielt den kompletten Fragebogen mit der Bezeichnung «Deepfake», während die andere Gruppe den identischen Fragebogen erhielt, aber durchweg mit der Bezeichnung «synthetische Medien» (inkl. Definition). Dieses erste Experiment sollte Aufschluss darüber geben, ob die Bezeichnung als «Deepfake» generell einen Einfluss auf die Risikowahrnehmung hat.

### 3.2.2. Hauptstudie

In der Hauptstudie wurden die zentralen Konzepte erhoben und ein zweites Experiment im September 2023 durchgeführt (n = 1361). Für das Onlineexperiment wurde rund der Hälfte der Teilnehmenden eine Literacy-Intervention, also eine kurze Hilfestellung zur Erkennung von Deepfakes, gezeigt. Wir untersuchten, ob diese Literacy-Intervention einen Effekt auf die Kompetenz zur Erkennung von Deepfakes hat. Zusätzlich wurde die Rolle von allgemeiner Medienkompetenz, Social-Media-Literacy und Internet-Skills gemessen und wie diese Faktoren mit der Fähigkeit, Deepfakes zu erkennen, korrelieren. Die Fähigkeit zur Erkennung wurde mit sechs unterschiedlichen Videos gemessen. Die verwendeten Videos stammen aus dem Unterhaltungsbereich (Humor, Witz) und dem Informations-

bereich (aktuelle Themen) und wurden ohne Ton abgespielt. Es wurden dazu bereits existierende reale Videos und Deepfakes recherchiert (jeweils drei) und im Pretest getestet. Um die möglichen Einflussfaktoren möglichst zu minimieren, wurden nur Videos mit Porträts von international sehr bekannten Personen verwendet. Es wurden Videos mit Tom Cruise, Elon Musk, Barack Obama, Hillary Clinton, Volodimir Selenskyj und Vladimir Putin verwendet.

Für das Experiment wurden die Teilnehmenden zufällig zwei Gruppen zugewiesen. Eine Gruppe erhielt eine Literacy-Intervention (Guess u.a. 2020; Potter 2013). Die zweite Gruppe erhielt keine zusätzlichen Hinweise zur Erkennung von Deepfakes. Die Literacy-Intervention beinhaltete eine kurze Anweisung, wie Deepfakes erkannt werden können. Ziel war es also zu testen, ob mit einer kurzen Intervention mit begrenztem Umfang ein positiver Effekt auf die Deepfake-Erkennungskompetenz erzielt werden kann. Die Literacy-Intervention wurde bewusst kurzgehalten und enthielt nur Hinweise, die für die gezeigten Videos hilfreich waren:

«Im folgenden Abschnitt geht es um die Beurteilung von Videos. Bevor Sie die Videos beurteilen, geben wir Ihnen einige Tipps, wie man Deepfakes erkennen kann. Bitte lesen Sie den Text auf dieser Seite genau durch. Nach 20 Sekunden erscheint die ‹Weiter›-Taste am Ende dieser Seite.

Folgende Punkte können Ihnen Hinweise geben, ob es sich bei einem Video um einen Deepfake handelt:

Die Bewegungen

- Die Augen der Person wirken fixiert und schauen nicht natürlich in die Kamera.
- Das Gesicht der Person scheint über dem Körper zu schweben, wenn sie sich bewegt.

Abgebildete Details

- Details wie Zähne, Haarsträhnen oder Ohrringe sehen nicht realistisch aus.
- Der Hintergrund ist verschwommen oder sieht unpassend aus.

Der Kontext

- Die Person sieht anders aus, als sie es sich gewohnt sind.
- Die Person verhält sich anders, als sie es erwarten würden.»

Die Videos (reale und Deepfakes) wurden als GIF umgesetzt, um Probleme beim Abspielen auf der Umfrageplattform bzw. den Geräten der Befragten zu vermeiden. Ob die GIFs auf den Geräten der Befragten funktionierten, wurde ganz am Anfang mit einer Animation einer Zahl getestet. Befragte, welche die Zahl nicht erkannten, wurden ausgeschlossen. Die Befragten mussten jedes der sechs Videos auf einer Skala von 1 bis 7 einordnen (Endpunkte «ganz sicher Deepfake» und «ganz sicher real»). Die Personen konnten das Video beliebig oft anschauen. Nach diesem Literacy-Test wurden die Personen zudem gefragt, wie gut sie die Person im Video oder das Video schon kennen.

Um die zentralen Forschungsfragen zu beantworten, wurden im Fragebogen vor der Literacy-Intervention und der Messung der Fähigkeit, Deepfakes zu erkennen, weitere Aspekte erhoben.

Es wurde erfasst, welche individuellen Erfahrungen Personen mit Deepfakes bereits gemacht haben. Wir fragten ab, ob sie Deepfakes gesehen haben, Deepfakes geteilt haben oder Deepfakes selbst hergestellt haben. Weiter fragten wir ab, auf welchen Kanälen (Social Media, Messengerdienste, journalistische Medien) und wie oft Menschen Deepfakes angetroffen haben.

Ein zentraler Frageblock beinhaltet die Wahrnehmung von Chancen und Risiken. Dazu stützen wir uns auf die Forschung zur Risiko- und Chancenwahrnehmung von Technologien (Siegrist/Visschers 2013) und künstlicher Intelligenz (Bao u.a. 2022). Wir fragen nach Risiken zu Politik, Medien, Wirtschaft und individuellen Aspekten sowie Chancen für die Wirtschaft und individuellen Aspekten.

Um *Medienkompetenz* (Ashley u.a. 2013; Maksl u.a. 2015), *Social-Media-Kompetenz* (Tandoc Jr u.a. 2021) und *Internet-Skills* (Hargittai/Hsieh 2012; Koc/Barut 2016) zu erheben, stützen wir uns auf etablierte und validierte Skalen.

Als weitere Variablen haben wir erhoben, welchen Effekt von Deepfakes die Befragten auf ihre eigene Meinung sowie auf die Meinung von anderen vermuten, den sog. Presumed Media Effect (Hong/Kim 2020; Tal-Or u.a. 2010). Weiter haben wir gemessen, wie die Befragten ihre eigene Kompetenz, Deepfakes erkennen zu können, bewerten. Dazu haben wir uns an die Forschung zur Erkennung von Desinformation orientiert (Corbu u.a. 2020). Schliesslich haben wir das Vertrauen in Politik, Medien und Wirtschaft erhoben (Slovic 1999).

## 3.3. Resultate

Im folgenden Kapitel werden die Resultate aus den beiden Teilstudien vorgestellt. Die Resultate sind nach den drei zentralen Forschungsfragen gegliedert. Zuerst werden Resultate zur allgemeinen Wahrnehmung von Deepfakes präsentiert. Im zweiten Teil steht die Wahrnehmung von Chancen und Risiken von Deepfake-Technologien in der Bevölkerung im Zentrum. In einem abschliessenden dritten Teil werden Resultate zur Deepfake-Erkennungskompetenz vorgestellt.

### 3.3.1. Erfahrungen mit Deepfakes

Unsere Resultate zeigen, dass die Schweizer Bevölkerung noch wenig Erfahrungen im Umgang mit Deepfakes hat. Nur etwas mehr als die Hälfte der Befragten (57 %) kennt den Begriff Deepfake, knapp die Hälfte (49 %) hat nach eigenen Angaben schon Deepfakes gesehen. Ein nicht unwesentlicher Teil der Befragten gab an, den Begriff Deepfake vor der Studie nicht gekannt, aber schon Deepfakes gesehen zu haben. Entweder waren ihnen Deepfakes unter einem anderen Namen bekannt oder unsere Studie hat sie zur Reflexion angeregt. Lediglich eine kleine Minderheit hat schon Deepfakes selbst hergestellt (2 %) oder im Netz weiterverbreitet (3 %).

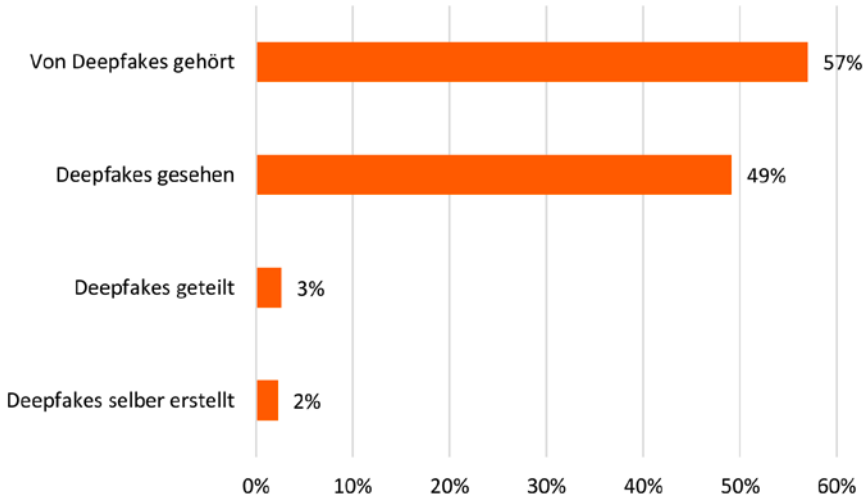


**Abbildung 13**: Erfahrungen mit Deepfakes (prozentualer Anteil der Befragten, die dem jeweiligen Item zugestimmt haben)

#### 3.3.1.1. Unterschiede zwischen soziodemografischen Gruppen

Die grundlegenden Befunde zu Erfahrungen mit Deepfakes unterscheiden sich nicht wesentlich nach Geschlecht, Alter, Bildung und Sprachregion. Fragt man nach der Bekanntheit des Begriffs Deepfakes, unterscheiden sich die Altersgruppen nur wenig. Etwas überraschend ist, dass ältere Menschen über 55 Jahren den Begriff eher kennen (63 %) als Personen unter 35 Jahren (52 %). Allerdings geben die Jüngeren (59 %) deutlich eher an, schon Deepfakes gesehen zu haben, als die Älteren (39 %). Männer (63 %) geben öfters als Frauen (54 %) an, den Begriff Deepfake vor der Studie schon gekannt zu haben. Bei der Frage, ob man schon Deepfakes gesehen hat, unterscheiden sich Männer (51 %) und Frauen (48 %) kaum. Personen mit Hochschulabschluss kennen den Begriff Deepfake eher (62 %) und haben schon Deepfakes gesehen (53 %) als Personen ohne Hochschulabschluss (55 % bzw. 49 %). In der Suisse romande (69 %) kennt man den Begriff Deepfake eher als in der Deutschschweiz (51 %). Allerdings haben die Deutschschweizer öfters Deepfakes gesehen (52 %) als Menschen in der Suisse romande (44 %). Über alle Gruppen hinweg hat nur eine Minderheit schon selbst Deepfakes geteilt oder erstellt. Am ehesten haben junge Männer unter 35 eigene Erfahrungen mit dem Erstellen (6 %) oder Teilen von Deepfakes (6 %).

Unsere Resultate zeigen insgesamt, dass die Menschen in der Schweiz eher wenig Erfahrungen mit Deepfake-Technologien haben. Typische Erklärungsfaktoren für die Adaption von Technologien, wie Alter, Geschlecht und Bildung, haben keinen allzu starken Einfluss. Es ist also nicht so, dass es in der Bevölkerung Segmente gibt, die viel Erfahrung mit Deepfakes haben, und andere mit sehr wenig Erfahrungen. Personen, die selbst Deepfakes kreieren, sind über alle soziodemografischen Gruppen hinweg aktuell noch eine kleine Minderheit.

#### 3.3.1.2. Medienkanäle und Deepfakes

Wir haben weiter gefragt, auf welchen Medienkanälen die Befragten auf Deepfakes treffen (1 = gar nie – 7 = sehr oft). Am ehesten treffen Schweizerinnen und Schweizer auf Social Media (M = 4,1) und Videoplattformen (4,0) auf Deepfakes. Deutlich weniger oft werden journalistische Medien (M = 3,4) oder Messengerdienste (M = 3,3) als Kanäle für den Kontakt mit Deepfakes angegeben. Unsere Studie zeigt also, dass Social Media wie Facebook, Instagram und TikTok sowie Videoplattformen wie YouTube oder Vimeo in der Schweiz als die primären Verbreitungskanäle für Deepfakes wahrgenommen werden. Dass Messengerdienste wie WhatsApp und Telegram eher wenig stark mit der Verbreitung von

Deepfakes in Verbindung gebracht werden, erstaunt insofern, da gerade diese Kanäle in der Schweiz stark mit der Verbreitung von Desinformation während der Coronapandemie assoziiert wurden. Allerdings haben wir nicht gefragt, welche Arten von Deepfakes auf den unterschiedlichen Kanälen angetroffen werden. Es können, müssen aber nicht potenziell schädliche Deepfakes sein.

Weiterführende explorative Analysen zeigen, dass die Nutzung von journalistischen Medien positiv mit der Bekanntheit des Begriffs Deepfakes korreliert ($b = 0.154$, $p \leq 0.001$). Die Nutzung von journalistischen Medien korreliert aber nicht mit dem Sehen von Deepfakes ($b = –0.040$, $p = 0.212$). Umgekehrt verhält es sich mit der Nutzung von Videoplattformen wie YouTube. Je häufiger man Videoplattformen nutzt, desto eher hat man schon Deepfakes gesehen ($b = 0.123$, $p \leq 0.001$). Eine Korrelation zwischen der Nutzung von Videoplattformen und der Bekanntheit des Begriffs Deepfake besteht hingegen nicht ($b = –0.015$, $p = 0.649$). Für Social Media und Messengerdienste liess sich in beiden Fällen keine Korrelation feststellen. Zusammenfassend lässt sich also sagen, dass die Nutzung von journalistischen Medien mit der Bekanntheit des Phänomens korreliert, die Nutzung von Videoplattformen hingegen mit direkten Erfahrungen mit Deepfakes.

### 3.3.2. Werden Deepfakes als Chance oder als Risiko wahrgenommen?

In einem weiteren Schritt haben wir gefragt, wie sehr sie Statements zum Risiko von Deepfake-Technologien für verschiedene Gesellschaftsbereiche einschätzen (1 = stimme überhaupt nicht zu bis 7 = stimme sehr stark zu). Deepfakes werden von Schweizerinnen und Schweizern stark mit Risiken assoziiert. Dabei stehen insbesondere Medien im Fokus, also die Angst, dass Deepfakes das Vertrauen in Schweizer Medien untergraben ($M = 5,5$) und für die Erzeugung von Falschnachrichten ($M = 6,1$) verwendet werden. Etwas weniger hoch werden die Risiken für die Politik eingeschätzt – konkret die Gefahr, dass Deepfakes in der Schweiz zur Manipulation von Wahlen ($M = 5,0$) oder Abstimmungen ($M = 5,0$) verwendet werden. Ähnlich hoch werden die Risiken für das Vertrauen in die Schweizer Wirtschaft ($M = 5,1$) und die Gefährdung des Wirtschaftsstandorts durch Deepfake-Technologien aus dem Ausland ($M = 4,7$) von den Befragten beurteilt. Individuelle Risiken durch Deepfakes, wie die Gefährdung der Privatsphäre ($M = 4,1$) oder die Gefahr, selbst Opfer eines Deepfakes zu werden ($M = 4,1$), schätzt die Schweizer Bevölkerung deutlich tiefer ein.

Wir haben zusätzlich für den Wirtschaftsbereich und individuelle Aspekte auch explizit nach Chancen von Deepfakes gefragt. Erstens muss festgehalten werden, dass die Chanceneinschätzungen (M = 3,1) auf einer Skala von 1 bis 7 insgesamt deutlich tiefer ausfallen als die Risikoeinschätzungen (M = 4,9). Die Schweizer Bevölkerung sieht also Risiken deutlich stärker im Vordergrund als allfällige Potenziale von Deepfake-Technologien. Chancen werden eher für den individuellen Bereich gesehen, namentlich dass Deepfakes die eigene Kreativität fördern (M = 3,3) und ein kurzweiliger Zeitvertreib (M = 3,3) sind. Chancen für neue Arbeitsplätze (M = 2,9) und Wettbewerbsvorteile von Deepfake-Technologien für den Schweizer Wirtschaftsstandort (M = 3,0) werden hingegen eher tief eingeschätzt.

#### 3.3.2.1. Mit Risikowahrnehmung korrelierende Faktoren

Weiterführende explorative Analysen zeigen, dass die Risikowahrnehmung kaum durch soziodemografische Faktoren bestimmt wird (vgl. Abbildung 14 bis Abbildung 17 sowie ausführlicher im Anhang Tabelle 18 bis Tabelle 21). Mit drei Ausnahmen: Erstens schätzen Personen aus der Suisse romande die Risiken von Deepfakes im Bereich von Politik (b = –0.46, 95 % CI [–0.62, –0.29], p ≤ 0.001) und Medien (b = –0.20, 95 % CI [–0.32, –0.07], p = 0.002) weniger hoch ein als Menschen, die in der Deutschschweiz wohnen. Zweitens schätzen Personen mit Hochschulabschluss die Risiken im Bereich der Wirtschaft (b = –0.18, 95 % CI [–0.34, -0.01], p = 0.033) sowie individuelle Risiken (b = –0.29, 95 % CI [–0.50, –0.08], p = 0.007) als weniger hoch ein als Personen ohne Hochschulabschluss. Drittens schätzen Frauen die Risiken im individuellen Bereich, also die Gefährdung der Privatsphäre oder selbst Opfer von Deepfakes zu werden, höher ein als Männer (b = 0.31, 95 % CI [0.11, 0.50], p = 0.002).

Auch der vermutete Effekt von Deepfakes auf die Schweizer Bevölkerung spielt bei der Risikobewertung eine Rolle. Je stärker eine Person davon ausgeht, dass Deepfakes einen Effekt auf die Meinung von Menschen in der Schweiz haben, desto eher werden Deepfakes als Risiko wahrgenommen. Das gilt für alle vier abgefragten Bereiche: Politik (b = 0.42, 95 % CI [0.32, 0.52], p ≤ 0.001), Medien (b = 0.30, 95 % CI [0.23, 0.38], p ≤ 0.001), Wirtschaft (b = 0.39, 95 % CI [0.30, 0.48], p ≤ 0.001) und individuelle Risiken (b = 0.17, 95 % CI [0.05, 0.29], p = 0.005). Mehr Vertrauen in Politik und Medien führt zu einer tieferen Wahrnehmung von Risiken von Deepfakes für die Politik (b = –0.08, 95 % CI [–0.15, –0.01], p = 0.023), aber auch für das Individuum (b = –0.10, 95 % CI [–0.18, –0.02], p = 0.015). Etwas überraschend ist der Befund, dass das Vertrauen mit höherer Risikowahrnehmung im Bereich Medien korreliert (b = 0.06, 95 % CI

[0.01, 0.12], p = 0.014). Weiter gilt, dass je stärker die wahrgenommene Verbreitung von Deepfakes ist, desto eher sehen die Befragten ein Risiko für Politik (b = 0.12, 95 % CI [0.05, 0.20], p = 0.002) und Medien (b = 0.18, 95 % CI [0.12, 0.23], p ≤ 0.001). Zudem korreliert die Einschätzung der Befragten, Deepfakes selbst erkennen zu können, mit einer tieferen Risikowahrnehmung für Medien (b = –0.07, 95 % CI [–0.10, –0.03], p = 0.001).

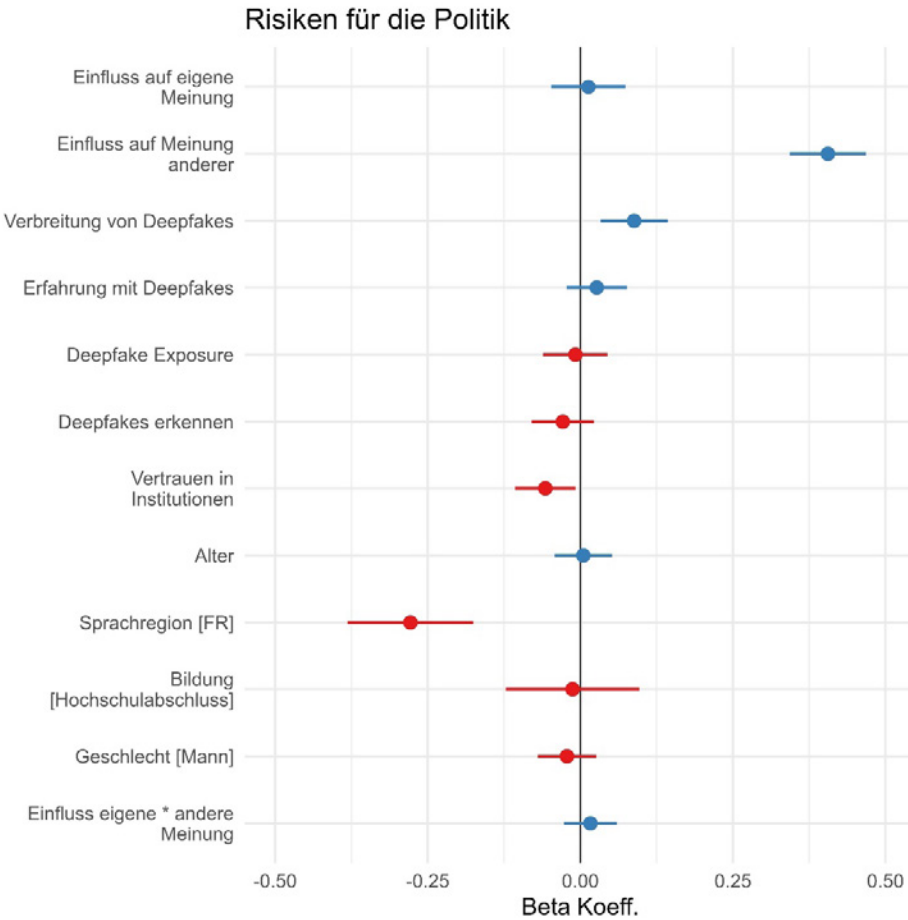


**Abbildung 14:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von Risiken von Deepfakes für die Politik. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Je höher der vermutete Einfluss von Deepfakes auf die Meinung von anderen, desto höher das wahrgenommene Risiko von Deepfakes für Medien.

Zur starken Risikowahrnehmung passt auch die Einstellung in der Bevölkerung bezüglich der Regulierung von Deepfake-Technologien. Die Befragten befürworten auf einer Skala von 1 (stimme überhaupt nicht zu) bis 7 (stimme voll und ganz zu) relativ klar alle Formen der Regulierung: sprich Gesetze (M = 5,2) und Verbote (M = 5,3) von Deepfake-Technologien sowie staatliche Regulierung (M = 5,0) und Selbstregulierung durch Internetunternehmen (M = 4,9).

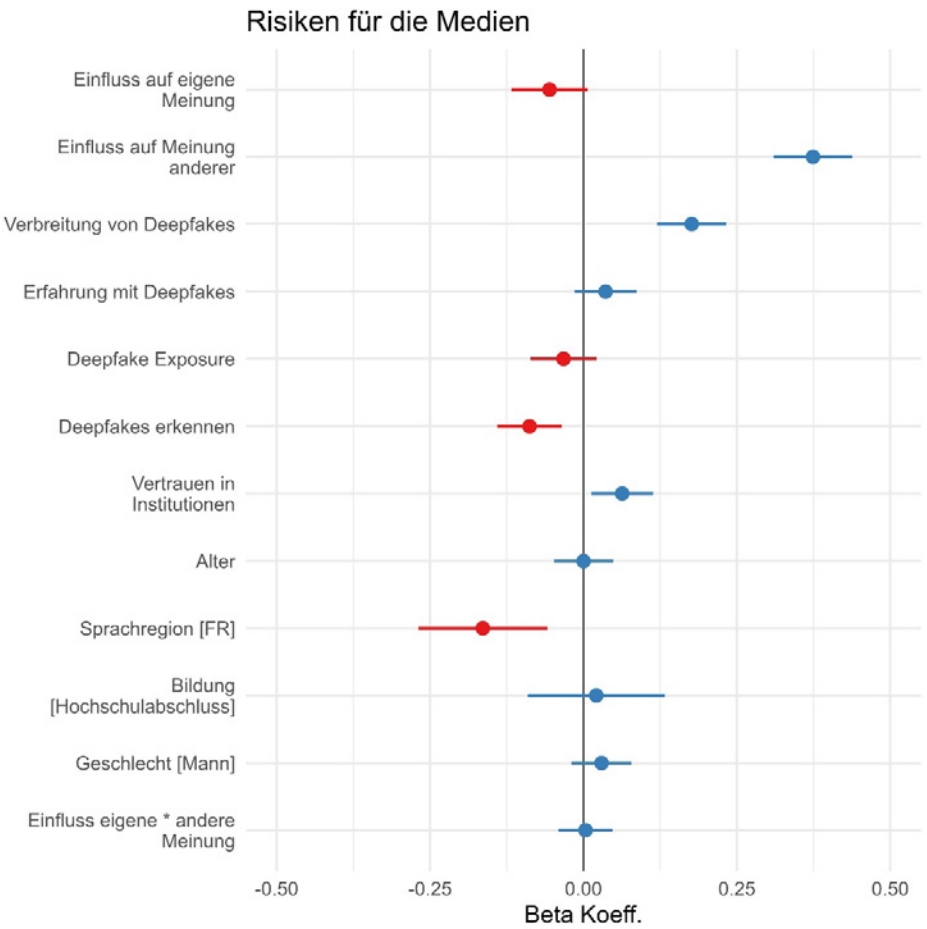


**Abbildung 15:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von Risiken von Deepfakes für die Medien. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Je höher die vermutete Verbreitung von Deepfakes, desto höher das wahrgenommene Risiko von Deepfakes für Medien.

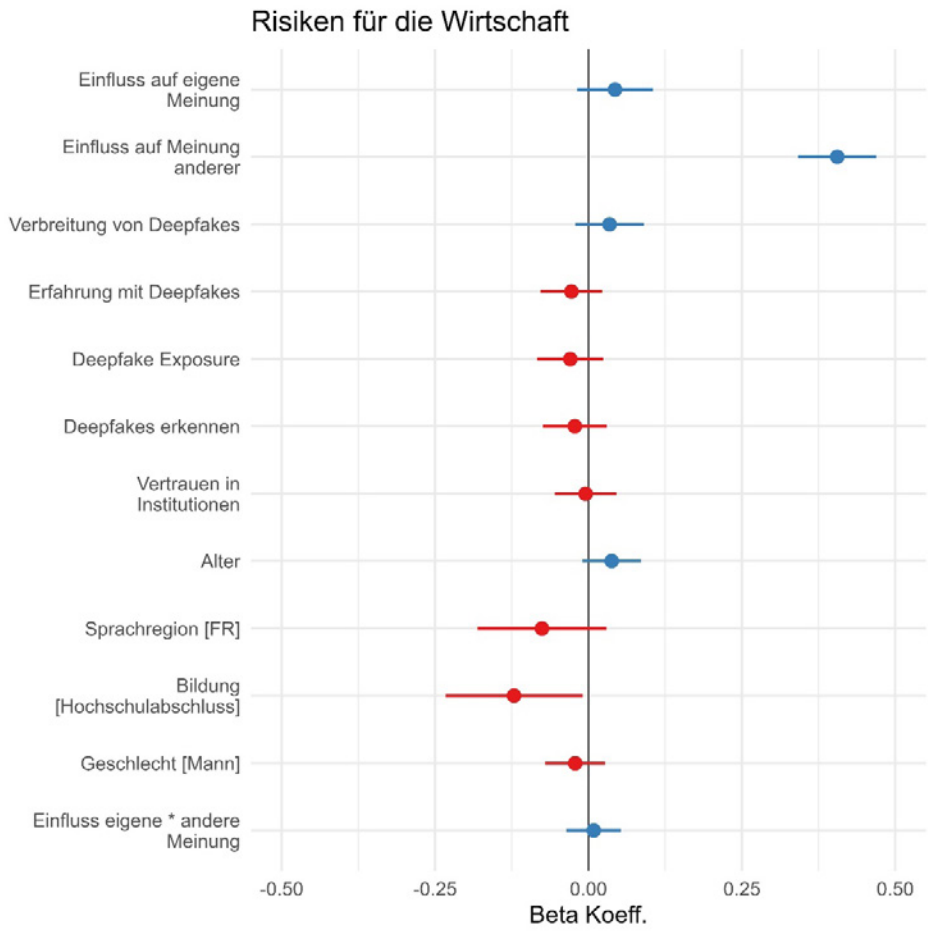


**Abbildung 16:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von Risiken von Deepfakes für die Wirtschaft. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Je höher der vermutete Einfluss von Deepfakes auf die Meinung von anderen, desto höher das wahrgenommene Risiko von Deepfakes für Medien.

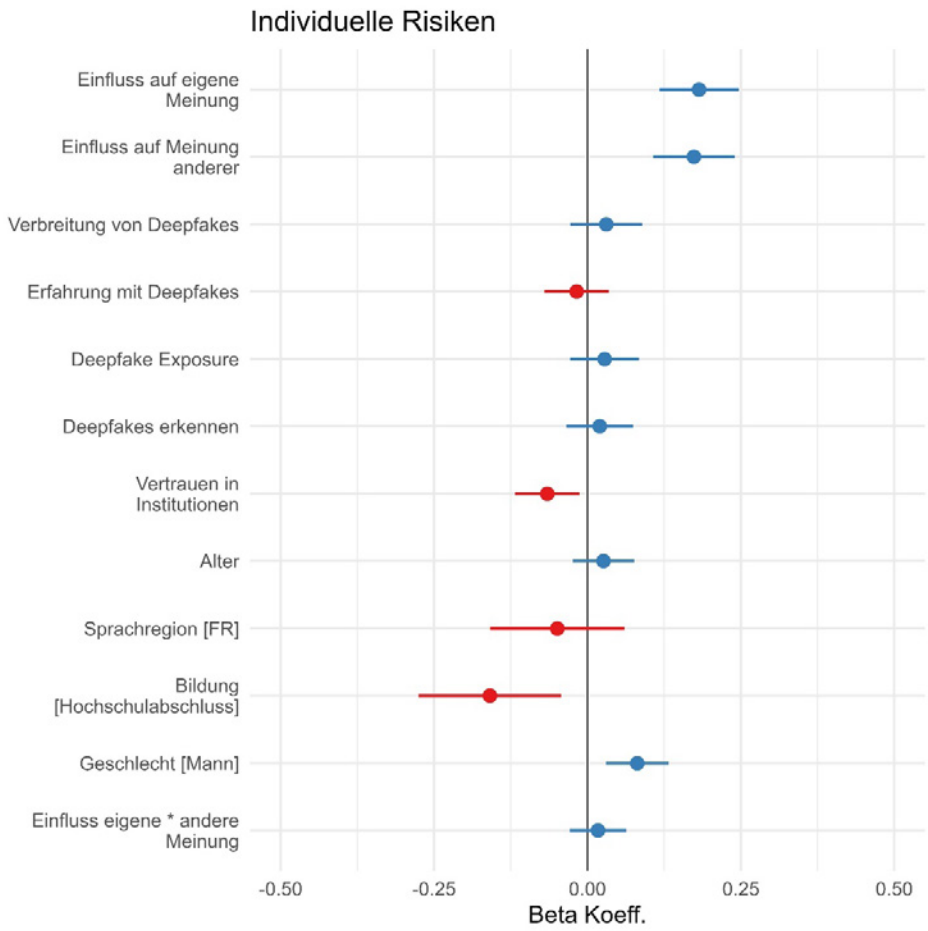


**Abbildung 17:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von individuellen Risiken von Deepfakes. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Frauen bewerten das individuelle Risiko von Deepfakes höher als Männer.

### 3.3.2.2. Einfluss des Labels (Deepfake vs. synthetische Medien)

In der ersten Teilstudie haben wir unterschiedliche Begriffe (Labels) verwendet, um die Risiken und Chancen von Deepfake-Technologien zu ermitteln. Eine Hälfte der Befragten erhielt den kompletten Fragebogen mit der Bezeichnung Deepfakes, die andere Hälfte mit dem Begriff synthetische Medien.

Zuerst zeigt sich, dass der Begriff synthetische Medien weniger bekannt ist. Wir haben, bevor wir die Definition gezeigt haben, ungestützt gefragt, was die Befragten mit dem Begriff Deepfake bzw. synthetische Medien assoziieren. Personen, die den Begriff Deepfake erhielten, gaben in 63 % der Fälle eine Antwort. Personen, die den Fragebogen mit der Bezeichnung synthetische Medien erhielten, nur in 38 % der Fälle. Offenbar weckt der Begriff Deepfakes in der Schweizer Bevölkerung eher Assoziationen als die Bezeichnung synthetische Medien.

Resultate zeigen weiter, dass die Risikobeurteilung von Deepfake-Technologien sich nicht signifikant vom Label (Deepfake vs. synthetische Medien) unterscheidet. Allerdings unterscheidet sich die Beurteilung von Chancen abhängig vom Label (vgl. Abbildung 18 bis Abbildung 20 sowie ausführlicher im Anhang Tabelle 22 bis Tabelle 24). Der Begriff synthetische Medien wird signifikant höher mit Chancen assoziiert als die Bezeichnung Deepfake. Das gilt für die abgefragten Dimensionen Medien ($b = -0.43$, 95 % CI [0.20, 0.66], $p \leq 0.001$) und Wirtschaft ($b = 0.31$, 95 % CI [0.16, 0.55], $p \leq 0.001$), nicht aber für individuelle Chancen ($b = 0.21$, 95 % CI [–0.01, 0.43], $p = 0.60$). Die Resultate zeigen, dass die Begrifflichkeit, mit der eine Technologie bezeichnet wird, mit entscheiden kann, wie Chancen und Risiken einer Technologie in der Bevölkerung eingeschätzt werden.

Neben dem Effekt des Labels zeigen sich einige signifikante Unterschiede zwischen soziodemografischen Gruppen. Je älter die Befragten, desto weniger sehen sie Chancen von Deepfake-Technologien für Medien ($b = –0.01$, 95 % CI [–0.02, –0.01], $p = 0.001$), Wirtschaft ($b = –0.01$, 95 % CI [-0.02, –0.01], $p \leq 0.001$) und individuelle Aspekte ($b = –0.02$, 95 % CI [–0.03, –0.01], $p = 0.001$). Das ist insofern interessant, weil Alter nicht mit der Beurteilung von Risiken von Deepfake-Technologien für verschiedene Gesellschaftsbereiche verbunden ist (vgl. Kapitel 3.3.2.1). Zudem sehen wir, dass Männer im Bereich der Medien eher Chancen sehen als Frauen ($b = 0.43$, 95 % CI –0.20, 0.66], $p \leq 0.001$). Im Bereich Wirtschaft bewerten Menschen aus der Suisse romande die Chance von Deepfake-Technologien tiefer als die Deutschschweizer ($b = –0.31$, 95 % CI [–0.54, -0.07], $p = 0.012$).

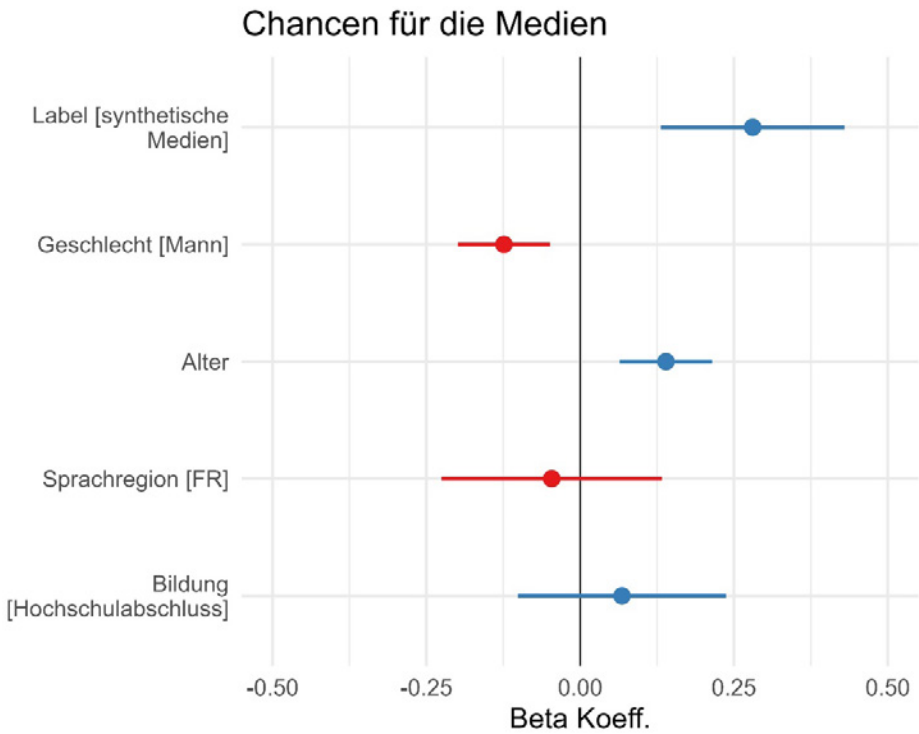


**Abbildung 18:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von Chancen von Deepfake-Technologien für Medien. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Die Bezeichnung (Label) von Deepfake als synthetische Medien führt zu einer höheren Einschätzung von Chancen von Deepfake-Technologien für die Medien.

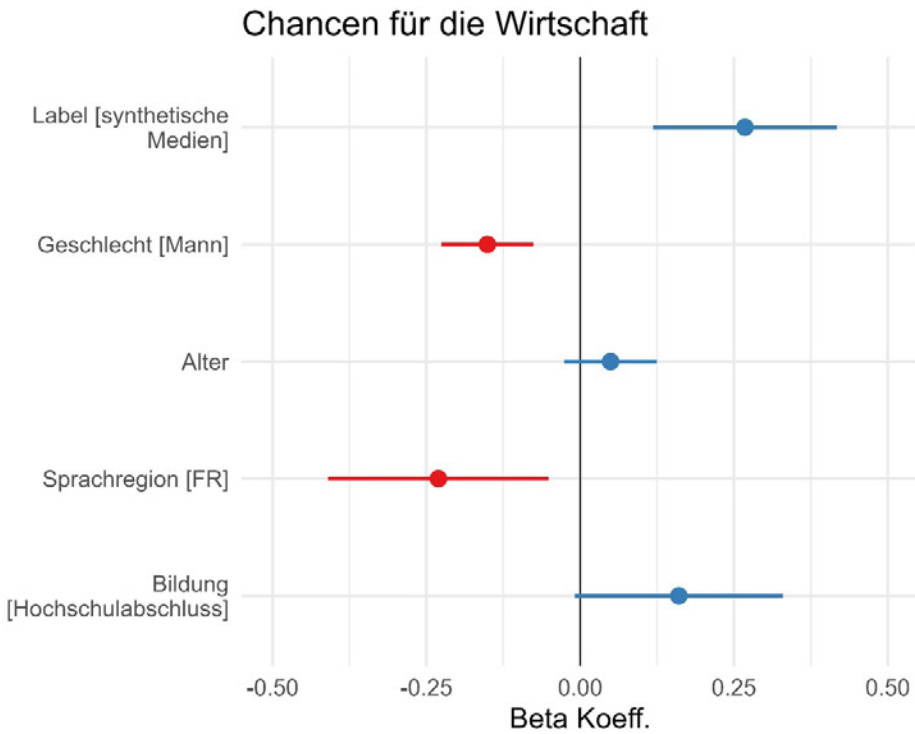


**Abbildung 19:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von Chancen von Deepfake-Technologien für die Wirtschaft. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Die Bezeichnung (Label) von Deepfake als synthetische Medien führt zu einer höheren Einschätzung von Chancen von Deepfake-Technologien für die Wirtschaft.

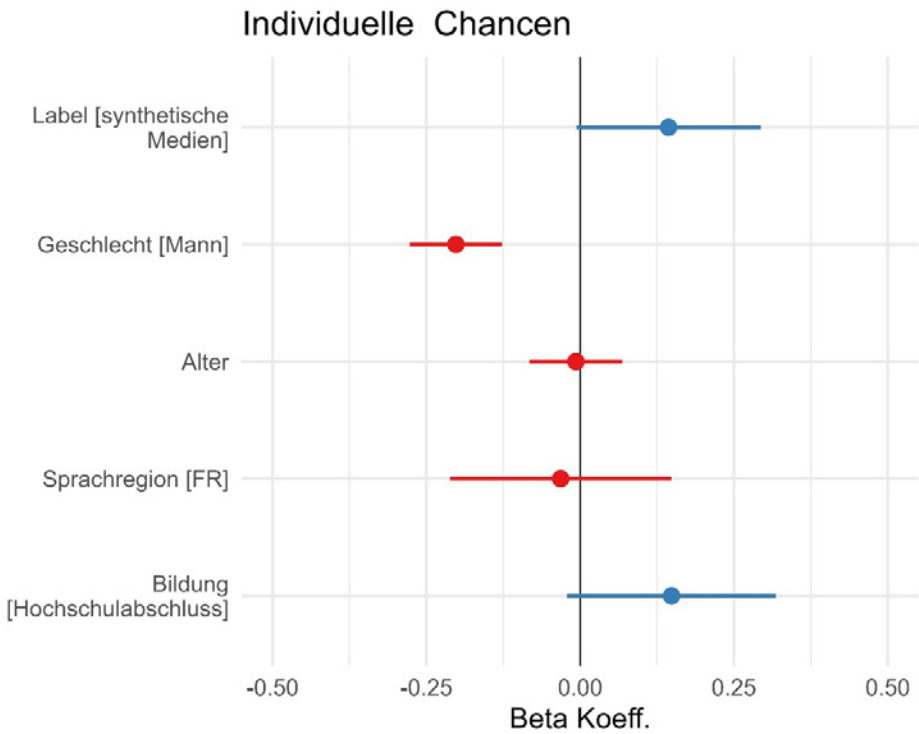


**Abbildung 20:** Beta-Koeffizienten mit 95%igem Konfidenzintervall zu Regressionsmodell für Wahrnehmung von individuellen Chancen von Deepfake-Technologie. Rot gefärbte Balken bedeuten ein negatives Beta, blau gefärbte Balken ein positives Beta. Balken, welche die Nullachse nicht schneiden, stehen für einen signifikanten Effekt der entsprechenden Variable (95 % Konfidenzintervall). Lesebeispiel: Frauen schätzen die individuellen Chancen von Deepfake-Technologien tiefer ein als Männer.

### 3.3.3. Kann die Schweizer Bevölkerung Deepfakes erkennen?

In einem letzten Teil haben wir getestet, ob Menschen in der Lage sind, Deepfakes von realen Videos zu unterscheiden. Wir haben den Befragten jeweils drei reale Videos und drei Deepfake-Videos gezeigt. Die Befragten mussten dann auf einer Skala von 1 (ganz sicher echt) bis 7 (ganz sicher ein Deepfake) einschätzen, ob das Video ein Deepfake oder ein reales Video war. Die Videos hatten drei unterschiedliche Schwierigkeitsgrade, die im Pretest ermittelt wurden.

Die Einschätzungen für das schwierige Beispiel eines Deepfake-Videos (M = 4,0) und eines realen Videos (M = 4,0) sowie des mittelschweren Deepfake-Videos (M = 4,1) lagen im Durchschnitt sehr nahe bei der Mittelkategorie von 4,0. Sprich, die Befragten waren sich sehr unsicher bei der Beurteilung der Videos. Bei den einfachen Beispielen für ein Deepfake- (M = 5,0) und ein reales Video (M = 3,7) sowie dem mittelschweren Beispiel für ein reales Video (M = 3,2) schnitten die Befragten beim Test ein wenig besser ab. Allerdings gilt insgesamt, dass die Befragten nur schwer Deepfake-Videos von realen Videos unterscheiden konnten, was sicher auch an dem relativ hohen Schwierigkeitsgrad der verwendeten Videos lag.

#### 3.3.3.1. Literacy-Intervention

Wir haben mit einem Modell getestet, ob eine kurze Hilfestellung, eine sog. Literacy-Intervention, die Deepfake-Erkennungskompetenz der Befragten verbessert (vgl. Abbildung 21). Unsere Daten zeigen keinen Effekt der Intervention (b = –0.06, 95 % CI [–0.23, 0.1], p = .451). Befragte, die eine Hilfestellung erhielten, haben also beim Erkennen von Deepfakes nicht besser abgeschlossen als Befragte, die keine Hilfestellung erhielten (vgl. Abbildung 21). Trotz des fehlenden positiven Effekts hatte die Literacy-Intervention auch keine negativen Auswirkungen. Unsere Daten zeigen keinen Backfire-Effekt: Eine Intervention führt also nicht zu einer überkritischen Beurteilung der Videos. Leute, die die Literacy-Intervention erhielten, schätzten reale Videos nicht eher als Deepfakes ein als Leute, die keine Literacy-Intervention erhalten hatten.

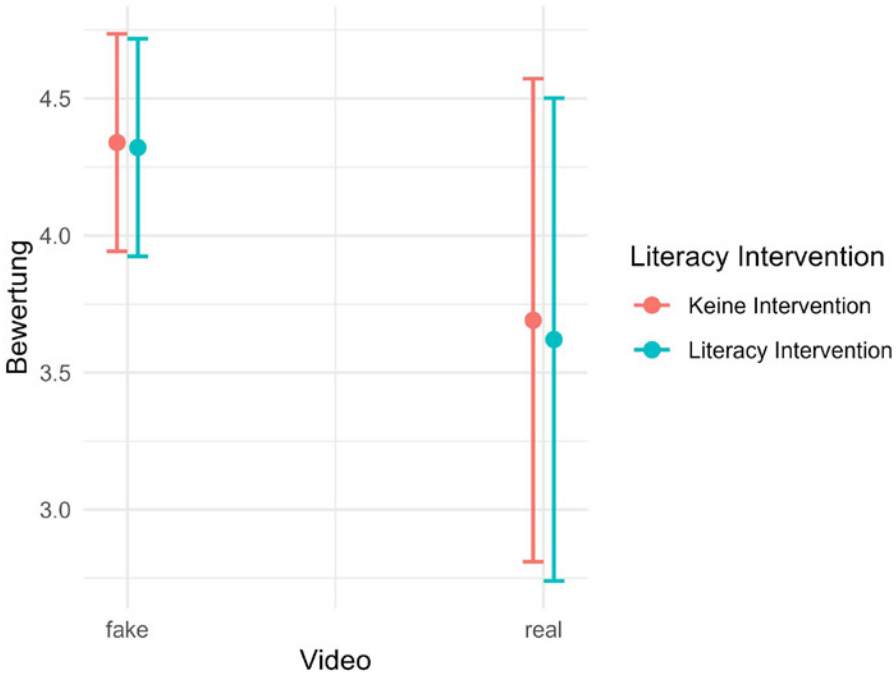


**Abbildung 21:** Effekt der Literacy-Intervention auf die Beurteilung von realen Videos und Deepfake-Videos. Auf der y-Achse ist die Bewertung von realen Videos und Deepfake-Videos auf einer Skala von 1 («ganz sicher echt») bis 7 («ganz sicher ein Deepfake») abgebildet, für Befragte, die eine Literacy Intervention (blau) erhalten haben, und solche, die keine Intervention erhalten haben (rot).

Zusätzlich haben wir noch explorativ untersucht, ob die Bekanntheit die korrekte Erkennung von Videos erklärt (vgl. Abbildung 22 sowie im Anhang Tabelle 25). Unsere Daten zeigen, dass dies der Fall ist (b = –0.07, 95 % CI [–0.11, –0.02], p = .003). Die Deepfake-Videos werden mit steigender Bekanntheit der dargestellten Person besser als Deepfake erkannt. Umgekehrt werden die realen Videos mit steigender Bekanntheit der Person im Video wahrscheinlicher als reale Videos eingestuft. Zudem hat Bildung einen positiven Effekt auf das korrekte Erkennen von Deepfakes (b = –0.20, 95 % CI [–0.39, –0.01], p = .037), wobei der Effekt stärker für die Erkennung von Deepfakes ist als für die Erkennung realer Videos.

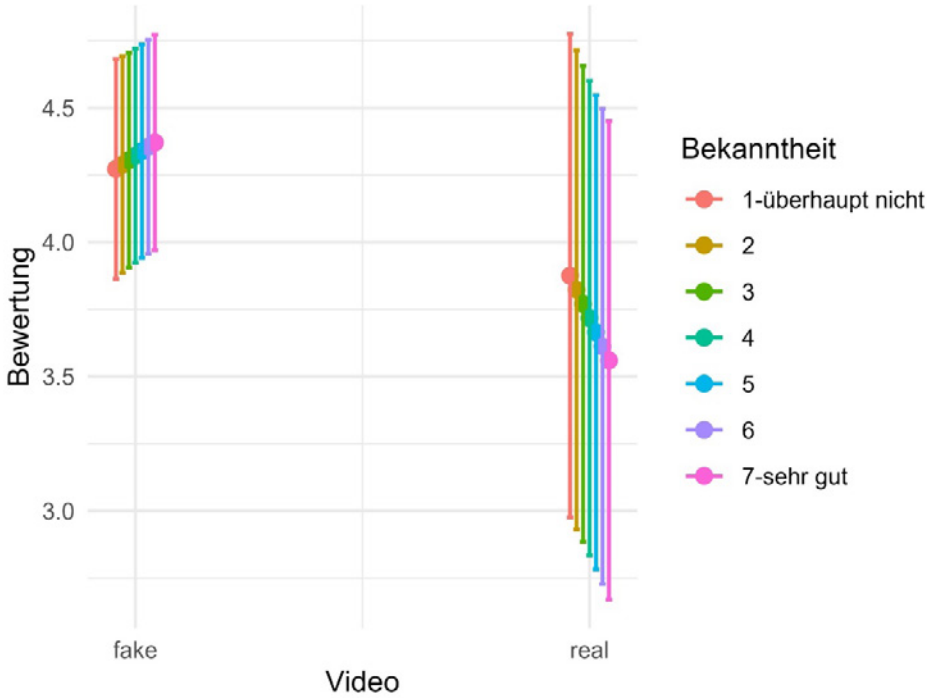


**Abbildung 22:** Einfluss von Bekanntheit auf die Beurteilung von realen Videos und Deepfake-Videos. Auf der y-Achse ist die Bewertung von realen Videos und Deepfake-Videos auf einer Skala von 1 («ganz sicher echt») bis 7 («ganz sicher ein Deepfake»), in Abhängigkeit von der durch die Befragten angegeben Bekanntheit des Videos, abgebildet.

### 3.3.3.2. Zusammenhang mit Medienkompetenz

Wir haben zusätzlich getestet, ob allgemeine Medienkompetenz, Social-Media-Literacy und Internet-Skills das korrekte Erkennen von Deepfakes positiv beeinflussen (vgl. Tabelle 26). In diesem Modell wurden nur die Teilnehmenden der Kontrollgruppe ohne Intervention analysiert. Die Ergebnisse bleiben allerdings gleich, wenn das gesamte Sample verwendet wird. Unsere Analyse zeigt nur für Social Media eine Korrelation (vgl. Abbildung 23 sowie im Anhang Tabelle 26). Je höher die Social-Media-Literacy, desto wahrscheinlicher werden Deepfakes und reale Videos richtig klassifiziert (b = −0.22, 95 % CI [−0.34, −0.11], p<.001).

Dabei ist der Zusammenhang stärker für die Deepfakes als für die realen Videos. Die anderen Literacy-Varianten spielen keine bedeutende Rolle. Das heisst, allgemeine Medienkompetenz und Internet-Skills korrelieren nicht mit der Fähigkeit, Deepfakes erkennen zu können. Soziodemografische Faktoren wie Alter, Geschlecht und Bildung sowie Erfahrung und Exposure zu Deepfakes haben ebenfalls keinen potenziellen Einfluss auf die Erkennungskompetenz.

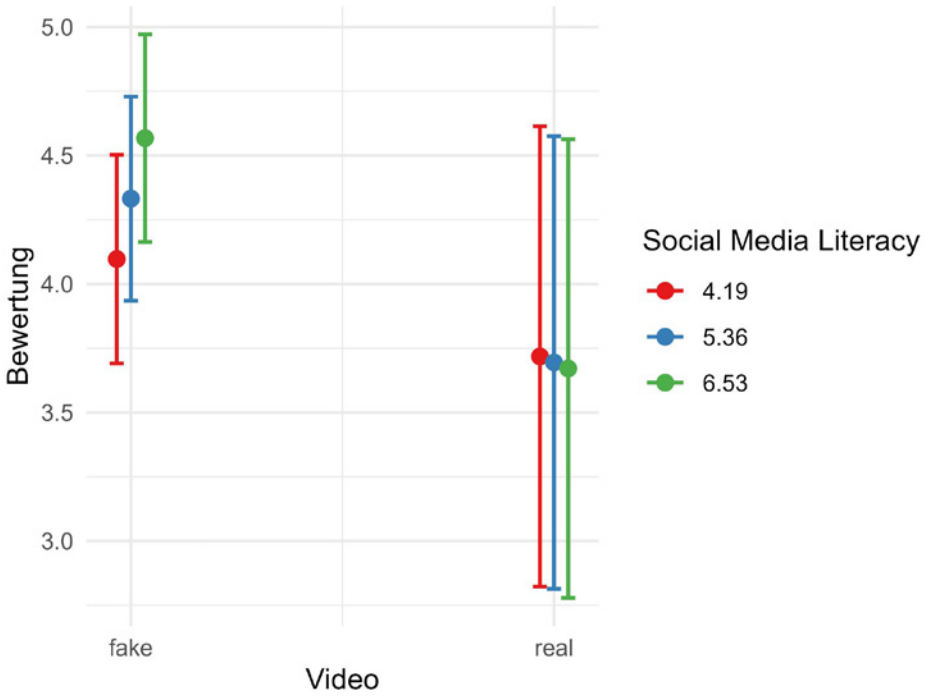


**Abbildung 23:** Einfluss von Social-Media-Literacy auf die Beurteilung von realen Videos und Deepfakes. Auf der y-Achse ist die Bewertung von realen Videos und Deepfake-Videos auf einer Skala von 1 («ganz sicher echt») bis 7 («ganz sicher ein Deepfake»), in Abhängigkeit von der Social-Media-Literacy der Befragten, abgebildet. Abgebildet ist der Mittelwert für Social-Media-Literacy (Skala 1–7) sowie jeweils plus und minus eine Standardabweichung.

## 3.4. Hauptbefunde

Deepfakes sind aktuell für viele Menschen ein noch unbekannter Begriff. Lediglich etwas mehr als die Hälfte der Befragten hat schon von Deepfakes gehört. Knapp die Hälfte hat nach eigenen Angaben bereits Deepfakes gesehen. Nur eine sehr kleine Minderheit hat schon Erfahrungen mit dem Erstellen und Verbreiten von Deepfakes. Das bedeutet einerseits, dass das Thema Deepfakes durch Unsicherheiten geprägt ist und viele Einschätzungen, Meinungen und Einstellungen zu Deepfakes auf Wahrnehmungen – beispielsweise über Diskurse in Medien – und nicht auf eigenen Erfahrungen beruhen. Andererseits stellt dies auch ein Opportunitätsfenster dar, da mit Bildungs- und Informationsangeboten zur Thematik ein sinnvoller Umgang mit Deepfake-Technologien erlernt werden kann.

Deepfakes werden von der Schweizer Bevölkerung stark mit Risiken assoziiert. Es gibt dabei kaum Unterschiede zwischen soziodemografischen Gruppen. Allerdings geht der vermutete Effekt von Deepfakes auf die Meinung von Drittpersonen mit einer höheren Risikowahrnehmung einher. Offenbar bestehen auch hinsichtlich Risiken eher diffuse Vorstellungen von Effekten von Deepfakes auf die Gesellschaft, welche die Risikowahrnehmung erhöhen. Die Verwendung der Bezeichnung synthetische Medien anstatt des Begriffs Deepfakes wirkt sich zudem positiv auf die Wahrnehmung von Chancen von Deepfake-Technologien aus. In der Summe deuten die Befunde darauf hin, dass noch viel Unsicherheit beim Thema Deepfake in der Bevölkerung besteht. Ein relevanter Befund ist zudem, dass Frauen individuelle Risiken – also die Gefährdung der Privatsphäre oder das Risiko, selbst Opfer eines Deepfake zu werden – höher einschätzen als Männer. In Anbetracht der Tatsache, dass viele bekannte Deepfake-Fälle aus dem pornografischen Bereich kommen, erstaunt dieses Resultat nicht. Allerdings müsste zukünftige Forschung untersuchen, welche Faktoren bzw. Erfahrungen für die höhere Risikowahrnehmung von Frauen tatsächlich verantwortlich sind.

Unsere Studie zeigt, dass Menschen in der Schweiz Deepfake-Videos kaum von realen Videos unterscheiden können. Wir haben für unser Experiment gut bis sehr gut gemachte Deepfake-Videos verwendet. Bei entsprechender Qualität lässt sich also ein Deepfake nur schwer erkennen. Umso wichtiger sind übergeordnete Medienkompetenzen. Unsere Studie zeigt, dass Social-Media-Literacy, also Kompetenzen im Umgang mit neuen Medien im Internet, positiv mit der Deepfake-Erkennungskompetenz korreliert. Zwischen dem Erkennen von Deepfakes und genereller Medienkompetenz sowie Internet-Skills gibt es hin-

gegen keinen Zusammenhang. Unsere kurze Literacy-Intervention, also eine kurze Hilfestellung zum Erkennen von Deepfakes unmittelbar vor dem Zeigen der Videos, hatte keinen Effekt auf die Erkennungskompetenz. Allerdings tritt auch kein Backfire-Effekt auf. Die Literacy-Intervention führt also nicht dazu, dass die Befragten überkritisch wurden und in der Folge auch reale Videos eher für Deepfakes hielten. Kurze Literacy-Interventionen könnten also auch dafür eingesetzt werden, um indirekt kritisches Denken zu aktivieren, und nicht zwingend, um direkte Effekte bei der Erkennung zu erzeugen. Diesen Pfad haben wir allerdings in der Studie nicht getestet. Weitere Forschung könnte hier zur Klärung beitragen.

# 4. Deepfakes im Recht

*Nula Frei & Sophia Rovelli*

Deepfakes werfen eine Reihe von juristischen Fragen auf. Neben der Frage, ob Deepfakes ggf. grundrechtlichen Schutz geniessen können (etwa unter der Meinungs- oder Kunstfreiheit) oder ob ein Urheberrecht an einem Deepfake bestehen kann (dazu nachfolgend Kapitel 4.1), stellt sich vor allem die Frage, welche Rechtsvorschriften durch Deepfakes verletzt werden und welche rechtlichen Möglichkeiten dagegen ergriffen werden können (nachfolgend Kapitel 4.2). Wie sich zeigen wird, existiert in der Schweiz derzeit keine spezifische Regulierung von Deepfakes, sondern es sind – je nach beabsichtigter Anwendung der Deepfake-Technologien – unterschiedliche allgemeine Rechtsvorschriften zu beachten. Über die zivil- und strafrechtlichen Verfahren hinaus gibt es auch öffentlich-rechtliche Vorgaben, namentlich im Rundfunkrecht sowie im Bereich der politischen Rechte, welche der Prävention missbräuchlicher Einsätze von Deepfakes dienen können (Kapitel 4.4). Schliesslich können Deepfakes auch im Rechtsalltag eine Rolle spielen, insbesondere als Beweismittel vor Gericht. Damit wird sich das Kapitel 4.3 auseinandersetzen. Zum Ende dieses Abschnitts wird Kapitel 4.5 mögliche Regulierungsansätze von Deepfakes darstellen.

## 4.1. Grundrechtlicher und urheberrechtlicher Schutz von Deepfakes

Deepfakes, die zum Beispiel zur Unterhaltung (Lantwin 2019; Meinicke 2020), als Kunstform oder im Bereich der Bildung verwendet werden, können unter Umständen einen eigenen rechtlichen Schutz beanspruchen. Dabei können sie entweder als künstlerische Ausdrucksform unter die Kunstfreiheit (Art. 21 BV[15]) fallen oder als Kundgebung bzw. Ausdruck einer Meinung oder einer Information in den Anwendungsbereich der Meinungs- und Informationsfreiheit (Art. 16 BV) fallen. Dieses Kapitel legt dar, inwiefern Deepfakes grundrechtlich durch die Meinungs-, Informations- oder Kunstfreiheit (Kapitel 4.1.1) sowie durch das Urheberrecht (Kapitel 4.1.2) geschützt sind.

---

[15] Bundesverfassung der Schweizerischen Eidgenossenschaft (BV, SR 101).

### 4.1.1. Meinungs-, Informations- und Kunstfreiheit

Der freie Austausch von Meinungen und Informationen bildet einen unentbehrlichen Bestandteil der rechtsstaatlichen Demokratie und ist Voraussetzung für eine pluralistische Gesellschaft (Hertig 2015a). Das Äussern und Erhalten von Informationen ist auch für den Einzelnen zur Persönlichkeitsentfaltung sowie zur Ausübung weiterer Freiheitsrechte von elementarer Bedeutung.[16]

Die Kommunikationsgrundrechte der Bundesverfassung[17] und Art. 10 der EMRK[18] gewährleisten den Schutz dieses Austausches.[19] Das Recht auf *Meinungs- und Informationsfreiheit* beinhaltet das Recht, sich eine eigene Meinung zu bilden, ungehindert zu äussern und zu verbreiten, sowie das Recht auf das freie Empfangen von Informationen anderer. Die Meinungsfreiheit umschliesst jegliche Ergebnisse von Denkvorgängen und mitgeteilten Überzeugungen und Äusserungen, unabhängig von deren Form (Hertig 2015a; 2015b).[20] Der Begriff der Meinung ist weit zu verstehen, folglich sind Inhalt oder Wertigkeit der Meinung im Prinzip unerheblich. Selbst unwahre, falsche oder irreführende Aussagen unterstehen bis zu einem gewissen Grad dem grundrechtlichen Schutz (Cueni 2019: 3 ff.).[21] So geht das herrschende Demokratieverständnis von mündigen Bürgerinnen und Bürgern aus, welche über ein kritisches Urteilsvermögen verfügen, einschliesslich der Fähigkeit, Übertreibungen und politische Des- und Misinformationen zu erkennen sowie zwischen unterschiedlichen Auffassungen zu differenzieren und vernünftige Entscheidungen zu treffen (Tschannen 2021: 398 N 1046).[22] Allerdings wird in der Rechtslehre postuliert, dass bewusst falsche oder irreführende Informationen – «Desinformation» also – nicht in den Schutzbereich fallen sollen. So ist auch das deutsche Bundesverfassungsgericht zum Schluss gekommen, dass «bewusst unwahre Tatsachenbehauptungen und solche, deren Unwahrheit bereits im Zeitpunkt der Äusserung unzweifelhaft

---

16 EGMR-Urteil vom 7.12.1976 «Handyside c. Grossbritannien».

17 Art.16, 17, 20, 21, 22, 23 BV.

18 (Europäische) Konvention zum Schutze der Menschenrechte und der Grundfreiheiten (EMRK, SR 0.101).

19 BGE 96 I 586 E. 6.

20 Art. 16 BV, siehe BSK-Kommentar, Hertig, Art. 16 BV N 9 ff.; Vorbemerkungen zu den Kommunikationsgrundrechten § 18 S. 213.

21 EGMR-Urteil vom 25.03.1985 «Barthold c. Germany».

22 BGE 98 Ia 73 E. 3b S. 80, BGE 117 Ia 41 E. 5a.

feststeht»,[23] nicht in den Schutzbereich der Meinungsfreiheit fallen. Aber auch wenn eine absichtliche Unwahrheit in den Schutzbereich eines Grundrechts fallen würde, würde die Tatsache, dass die Information in bewusst täuschender oder irreführender Absicht verbreitet wurde, im Rahmen der Interessenabwägung berücksichtigt und würde dazu führen, dass das öffentliche Interesse an der Einschränkung dieser Unwahrheit (z.B. ein Verbot ihrer Verbreitung) höher gewichtet würde (Rechsteiner/Errass 2014).

Die *Kunstfreiheit* schützt jegliche Art von künstlerischem Ausdruck.[24] Auch hier ist weder der Inhalt der Aussage, ihre Form, noch ihre Qualität für die Schutzwürdigkeit ausschlaggebend (Wyttenbach 2015). Hingegen ist es umstritten, ob auch vollautomatisch, also ohne menschliches Zutun geschaffene Werke von der Kunstfreiheit geschützt sind. Klar ist zwar, dass die Software (die «KI») selber nicht Grundrechtsträgerin sein kann, denn dies können nur natürliche oder juristische Personen sein (ebd.: N 5). Der persönliche Schutzbereich des Grundrechts kann sich also nur auf die den Algorithmus nutzenden Personen, z.B. die Urheberin oder den Verbreiter des Deepfakes, erstrecken. Die Frage ist aber, ob ein Deepfake als Werk in den sachlichen Schutzbereich der Kunstfreiheit fällt, ob also die Person, die einen Deepfake erstellt hat, sich darauf berufen kann, dass es sich hierbei um Kunst handle. Was Kunst ist (und was nicht), ist naturgemäss schwer zu definieren. Die Rechtsprechung behilft sich mit der Einteilung in einen materiellen (inhaltlichen) und formellen Kunstbegriff (ebd.: N 6). So betont etwa das deutsche Bundesverfassungsgericht, dass das «Wesentliche jeder künstlerischen Betätigung die freie schöpferische Gestaltung sei, in der Eindrücke, Erfahrungen, Erlebnisse des Künstlers ihren Ausdruck fänden». Künstlerische Tätigkeit sei «ein Ineinander von bewussten und unbewussten Vorgängen und Ausdruck der individuellen Persönlichkeit des Künstlers».[25] Nach einer anderen Umschreibung ist die Kunstfreiheit anwendbar, wenn «ein kreativer, schöpferischer Akt vorliegt – eine Verbindung aus Phantasie, Intellekt und Technik – und eine gewisse, wie auch immer ausgestaltete Kenntnisnahme durch die Aussenwelt erfolgt» (Wyttenbach 2015: N.6). So gesehen kann es nicht die Software sein, auf die entscheidend beim künstlerischen Ausdruck abzustellen ist. Weder besitzt ein Computeralgorithmus eine Persönlichkeit, noch ein Bewusstsein oder eine Gefühlswelt, die er zum Ausdruck bringen könnte

[23] BVerfG 99, 185, 197 – Scientology. Diese Meinung vertritt in Bezug auf Deepfakes auch Linardato (2021).

[24] Art. 21 BV.

[25] BVerfG 30, 173, 188 – Mephisto.

(Lewke 2017: 207 ff.). Andere Lehrmeinungen vertreten jedoch, dass auch die Kunstwerke einer subsymbolisch trainierten künstlichen Intelligenz der Ausdruck schöpferischer Tätigkeit sein können, indem sie – auf ihre Weise in Form von Feedbackschleifen – Eindrücke, Erfahrungen und Erlebnisse zu einem Kunstwerk verdichten und insofern als ein Ineinander von bewussten und unbewussten Vorgängen zu verstehen sind, die nicht rational aufzulösen sind (Kersten 2020). Zu dieser Frage besteht also noch Klärungsbedarf in Rechtslehre und Rechtsprechung.

Das Recht auf *Medienfreiheit* schliesslich garantiert die öffentliche fernmeldetechnische Verbreitung von Darbietungen und Informationen.[26] Nebst traditionellen Medien wird gleichermassen die Kommunikation über das Internet geschützt (Nobel/Weber 2021: 67).

Die Schutzbereiche der Kommunikationsgrundrechte sind also weit gefasst. Entsprechend der hier vertretenen Ansicht dürften Deepfakes grundsätzlich in den sachlichen Schutzbereich der Meinungs- und Informations- und unter Umständen auch der Kunstfreiheit fallen, allerdings ist der Natur ihrer Entstehung, ihrer potenziell schädlichen Auswirkungen sowie der damit verfolgten Absichten im Rahmen der Interessenabwägung Rechnung zu tragen, sodass eine staatliche Einschränkung der Erstellung oder Verbreitung von Deepfakes regelmässig gerechtfertigt sein dürfte. Zu beachten ist, dass die Grundrechtsträgerschaft sich auf natürliche und juristische Personen beschränkt. Ein Algorithmus ist selber nicht Grundrechtsträger; grundrechtlichen Schutz kann nur diejenige natürliche (oder u.U. juristische[27]) Person beanspruchen, die den Deepfake erstellt resp. verbreitet.

Entsprechend stellt eine staatliche Regulierung von Deepfakes, welche deren Erstellung und Verbreitung limitiert, einen Grundrechtseingriff dar, welcher den Vorgaben an Grundrechtseinschränkungen in Art. 36 BV (öffentliches Interesse, Verhältnismässigkeit, Wahrung des Kerngehalts) unterliegt.

In Bezug auf Deepfakes bedeutet dies, dass sie als Ausdrucksformen, die in den Schutzbereich der Kommunikationsgrundrechte fallen, zwar geschützt sind, aber rechtmässig eingeschränkt werden können, namentlich mit Blick auf öffentliche Interessen wie die öffentliche Sicherheit, der Schutz von Grundrechten

---

[26] Art. 17 URG.

[27] Juristische Personen können ebenfalls Grundrechtsträger sein, zu denken ist insb. an Medienhäuser im Rahmen der Meinungs- sowie der Medienfreiheit.

Dritter, der Schutz des Ansehens und der Unparteilichkeit der Justiz und weitere. Eine systematische Zensur vor der Veröffentlichung von jeglichen Inhalten im Internet mit Ziel einer Verhinderung von Verbreitung von Deepfakes wäre nicht zulässig. Die bestehenden Gesetze (RTVG[28], StGB[29], ZGB[30], DSG[31]), die zum Schutz von öffentlichen Interessen oder Grundrechtschutz Dritter Kommunikationsgrundrechte einschränken, sind auch bei der Erstellung und Verbreitung von Deepfakes zu beachten (und werden in den folgenden Unterkapiteln behandelt).

### 4.1.2. Urheberrecht

Deepfakes werden in Form von Video-, Foto- oder Audiodateien verbreitet. Es ist zu prüfen, ob sie als Kunstwerke oder Tonherstellungen im Sinne des Urheberrechts gelten und sie sich somit urheberrechtlich *schützen* lassen. Die Frage wiederum, ob es beim Erstellen von Deepfakes und deren anschliessender Verbreitung zu einer Urheberrechts*verletzung* kommt – etwa wenn fremdes Bildmaterial verwendet wird –, wird in einem späteren Kapitel (vgl. 4.2.1.3) analysiert.

Das Bundesgesetz über das Urheberrecht und verwandte Schutzrechte (URG)[32] regelt den nationalen Schutz von Urheberinnen von Werken der Literatur, der Kunst sowie Herstellerinnen von Ton. In unserem Kontext sind allerdings häufig mehrere nationale Rechtsordnungen anwendbar, da ein Deepfake über das Internet in mehreren Ländern erstellt und verbreitet wird. Die Frage, welche Rechtsordnung angewandt wird, ist von Bedeutung, da der Schutzumfang unterschiedlich ausgestaltet sein kann (Cappello 2020). Die revidierte Berner Übereinkunft zum Schutz von Werken der Literatur und Kunst[33] sorgt in den beteiligten Vertragsstaaten allerdings für einen vergleichbaren Urheberrechtsschutz. Die Frage des einschlägigen nationalen Rechts ist gemäss Art. 110 Abs. 2 IPRG[34] zu ermitteln. Die folgenden Ausführungen beschränken sich auf die Schweizer Rechtsordnung.

---

[28] Bundesgesetz über Radio und Fernsehen (RTVG, SR 784.40).

[29] Schweizerisches Strafgesetzbuch (StGB, SR 311.0).

[30] Schweizerisches Zivilgesetzbuch (ZGB, SR 210).

[31] Bundesgesetz über den Datenschutz (DSG, SR 235.1).

[32] Bundesgesetz über das Urheberrecht und verwandte Schutzrechte (URG, SR 231.1).

[33] SR 0.231.15.

[34] Bundesgesetz über das internationale Privatrecht (IPRG, SR 291).

Werke sind nach Schweizer Recht unabhängig von deren Wert oder Zweck vom Urheberrechtsschutz umfasst. Der Schutz des Werks beginnt zum Zeitpunkt der Schöpfung und die Schutzdauer beträgt 50 oder 70 Jahre. Dem Urheber eines Werks kommt das ausschliessliche Recht am eigenen Werk zu und er kann damit bestimmen, ob und wie sein Werk verwendet wird. Schutzwürdig können unter anderem auch fotografische, filmische, sonstige visuelle oder audiovisuelle Werke sein. Ob ein Werk gemäss Urhebergesetz vorliegt, ist im Einzelfall auszulegen. Eine geistige Schöpfung mit individuellem Charakter wird zur Erfüllung des Werkbegriffs grundsätzlich vorausgesetzt.

Damit das Werk das Erfordernis der geistigen Schöpfung erfüllt, muss es auf einem menschlichen Willen beruhen und eine Gedankenäusserung ausdrücken.[35] Ein Werk, welches von einem Computer allein generiert wird, erlangt folglich keinen Schutz. Jedoch gilt eine Kreation, welche von einem Menschen mithilfe eines Computers erschaffen wird, als urheberrechtliches Werk (Nobel/Weber 2021: 711).[36] Der für das Vorliegen eines Werks erforderliche individuelle Charakter wird in der Rechtslehre als statistische Einmaligkeit beschrieben. Der künstlerische oder ästhetische Wert eines Werks ist für die Festlegung des individuellen Charakters unbedeutend (ebd.: 712). Ferner sind an die Originalität keine hohen Anforderungen zu stellen.[37] Für Fotografien, die ein dreidimensionales Objekt wiedergeben, besteht zudem eine gesetzliche Ausnahme. Diese erfüllen den Werkbegriff, ungeachtet davon, ob ihnen ein individueller Charakter zugeschrieben werden kann (Art. 2 Abs. 3 URG).

Es stellt sich die Frage, ob Anwendungen von Deepfakes den Werkbegriff des Urhebergesetzes erfüllen. Einer «künstlichen Intelligenz» alleine kann kein Urheberrechtsschutz an einem Werk zugesprochen werden. Erstellt ein kommerziell verfügbares Programm automatisiert ein Deepfake, entsteht daraus keine geistige Schöpfung. Die blosse menschliche Mitarbeit im Sinne des Auslösens des Trainingsprozesses wird nicht ausreichen, damit eine geistige Schöpfung gemäss URG vorliegt (Beranek Zanon 2022). Anders verhält es sich unseres Erachtens, wenn eine kreative Tätigkeit der menschlichen Herstellerin beigesteuert wird, beispielsweise durch nachträgliches Anpassen oder durch die Bilderauswahl. Der Vorgang könnte hier mit dem Verwenden eines Fotoapparates durch einen Menschen gleichgestellt werden (AAP 2020). Der notwendige individuelle Charakter wäre im Einzelfall abzuklären, aber in letztgenannten Fällen

[35] BGE 130 III 168 S. 173, BBl 1989 III 521.

[36] BGE 130 III 173; BGE 116 II 352 ff.

[37] BGE 110 IV 102, 105.

wohl regelmässig zu bejahen (Beranek Zanon 2022), wobei hierzu unterschiedliche Auffassungen herrschen, gerade wenn der Zweck des Deepfakes in einer Imitation oder Täuschung besteht (ebd.). Je nach Herstellungsmethode dürften sich diese Fragen im Einzelfall aber auch anders beurteilen lassen, weshalb ein gewisser Interpretationsspielraum besteht. Schliesslich könnte auch an der Software selbst – welche die Deepfakes erstellt – ein urheberrechtlicher Schutz geltend gemacht werden (AAP 2020; Barrelet u.a. 2020).

## 4.2. Schutz vor schädlichen Auswirkungen von Deepfakes

Im Folgenden wird geprüft, welche Rechtsvorschriften gegen die Erstellung oder Verbreitung von schädlichen Deepfakes ins Spiel gebracht werden können. Vorbemerkt sei noch einmal, dass selbstverständlich nicht jeder Deepfake rechtswidrig ist. Die Erstellung synthetischer Medien ist grundsätzlich zulässig, solange sie keine Rechte anderer verletzen. Die nachfolgenden Ausführungen konzentrieren sich deshalb auf problematische Anwendungen von Deepfakes, also insbesondere solche, die in täuschender, manipulierender, rufschädigender oder sonst wie schädlicher Art oder Absicht erstellt bzw. verbreitet werden.

Im Folgenden wird zunächst auf die zivilrechtlichen Fragen eingegangen (Kapitel 4.2.1) und daraufhin auf das Strafrecht als Hebel gegen schädliche Deepfakes (Kapitel 4.2.2). Im Anschluss wird die Möglichkeit der verfahrensrechtlichen Geltendmachung allfälliger Abwehransprüche im Zivil- und Strafverfahren untersucht (Kapitel 4.2.3).

### 4.2.1. Schutz im Zivilrecht

#### 4.2.1.1. Persönlichkeitsverletzungen durch Deepfakes

Eine grosse Mehrheit der kursierenden Deepfakes bildet Personen ohne deren Einverständnis ab. Deshalb drängt sich die Frage auf, ob eine widerrechtliche Persönlichkeitsverletzung vorliegt. Der Persönlichkeitsschutz ist in den Art. 27 ff. ZGB festgehalten. Ein Eingriff in die Persönlichkeit ist widerrechtlich, sofern kein Rechtfertigungsgrund vorgebracht werden kann. Das Vorliegen wird in einem zweistufigen Verfahren geprüft: 1. Liegt eine Persönlichkeitsverletzung vor? 2. Sind allfällige Rechtfertigungsgründe gegeben? (Nobel/Weber 2021: 280 f.;

Melli 2022: N 39)[38] Der Persönlichkeitsschutz umfasst verschiedene Persönlichkeitsgüter, dazu gehören die Ehre, die Privatsphäre, das Recht am eigenen Bild, das Recht an der eigenen Stimme sowie das Recht am eigenen Namen.

- Der *Schutz der Ehre* einer Person bezieht sich auf den Ruf der Person in der Öffentlichkeit sowie das persönliche, nach innen gerichtete Ehrgefühl (Nobel/Weber 2021: 289). Durch diese Bestimmung erfasst wird das berufliche, wirtschaftliche und das gesellschaftliche Ansehen einer Person (Melli 2022: N 28).[39] Die Ehre kann auch ohne Namensnennung verletzt werden, wenn aufgrund der Umstände die Identität der Person für Dritte erkennbar wird (Nobel/Weber 2021: 295).[40] Das Ansehen einer Person kann auch durch Bilder, Fotomontagen, Karikaturen oder Videos herabgesetzt werden. Ehrverletzungen können als Tatsachenbehauptungen oder Werturteile geäussert werden. Tatsachenbehauptungen beziehen sich auf konkrete Ereignisse. Diese sind persönlichkeitsverletzend, falls sie nicht der Wahrheit entsprechen oder wenn sie aufgrund Ungenauigkeit, Verallgemeinerung oder Verkürzung persönlichkeitsverletzend wirken.[41] Ein Werturteil ist Ausdruck von Geringschätzung oder Missachtung einer Person. Werturteile sind zulässig, sofern sie aufgrund des Sachverhaltes vertretbar sind oder die Ehre an sich nicht streitig machen.[42] Der Beurteilungsmassstab einer Ehrverletzung bezieht sich nicht auf das subjektive Empfinden der betroffenen Person, sondern objektiviert sich anhand der Würdigung der konkreten Umstände (Nobel/Weber 2021: 292).[43] Der Schutz der Persönlichkeit gilt auch im Internet.
- Das *Recht auf Privatsphäre* ist gemäss Art. 28 ZGB geschützt. Es beinhaltet das Selbstbestimmungsrecht der Person über die Wiedergabe der sie betreffenden Informationen (Nobel/Weber 2021: 302; Hausheer/Aebi-Müller 2020).[44] Der Schutzumfang beschränkt sich auf die Informationenweiterga-

[38] BGE 136 III 410 E. 2.2.1; Art. 28 N 39.

[39] Der Schutz geht damit weiter als das Strafrecht, siehe unten 4.2.2.

[40] BGE 55 II 94 E. 1.

[41] BGE 129 III 48 E. 2.2.

[42] BGE 126 III 305 E. 4b/bb.

[43] BGE 129 III 35; BGE 132 III 641; BGE 138 III 641.

[44] BGE 143 III 297 E. 6.4.2.

be, durch welche die Person tatsächlich und spürbar beeinträchtigt wird.[45] Die Privatsphäre wird traditionell in drei unterschiedlich sensible Sphären eingeteilt: Geheim-/Intimsphäre, Privatsphäre und die Gemein-/Öffentlichkeitssphäre (Aebi-Müller 2005; Melli 2022: N 23).[46] In die Geheim-/Intimsphäre gehören diejenigen Lebensvorgänge, die eine Person der Wahrnehmung und dem Wissen aller Mitmenschen entziehen resp. nur bestimmten Personen mitteilen möchte.[47] Für öffentlich bekannte Personen ist der Schutzbereich der Geheimsphäre enger gefasst. Ob die Privatsphäre verletzt wird, ist nur anhand konkreter Umstände auszumachen (Nobel/Weber 2021: 302 f.).

- Das *Recht am eigenen Bild* stellt einen Abwehranspruch gegen das gezielte – auf Identifikation ausgerichtet – Erstellen von Abbildungen dar. Es beinhaltet auch das Recht auf Selbstbestimmung der Betroffenen bezüglich der Veröffentlichung des eigenen Bildes (Nobel/Weber 2021: 310 f.).[48] Auch die unbefugte Verwendung und Weiterverwendung von rechtmässig erstellten oder veröffentlichten Bildern verletzt das Recht am eigenen Bild (ebd.: 312).
- Das *Recht auf die eigene Stimme* und das *Recht am eigenen Wort* schützen vor unerlaubter Beschaffung, Verbreitung und Veränderung von Tonaufnahmen. Im Weiteren wird die Verwendung falscher Zitate und der Abdruck erfundener Interviews erfasst (Melli 2022: N 22).
- Art. 29 ZGB sieht den *Schutz des eigenen Namens* vor. Diese Bestimmung wird verletzt, wenn jemand sich den Namen einer anderen Person anmasst.

Die meisten Deepfakes (Schätzungen gehen von bis zu 95 % aus; van Huijstee u.a. 2021) betreffen pornografische Darstellungen von Frauen. Weitere denkbare Deepfake-Anwendungen sind Manipulationen von Videos von Politikern, die diese bei ehrverletzenden oder strafbaren Tätigkeiten zeigen sollen.[49]

Wird mithilfe technischer Mittel eine Nachbildung einer Person erstellt, erfolgt eine Persönlichkeitsverletzung. Das Recht auf Selbstbestimmung der eigenen Darstellung wird dadurch missachtet. Die Verwendung eines Abbilds einer Person in einem nicht vorgesehenen Zusammenhang stellt eine Persönlichkeitsver-

---

[45] BGE 143 III 297 E. 6.4.2.

[46] BGE 97 II 100 f.; 109 II 357; BGE 118 IV 45; 119 II 222 ff.

[47] BGE 118 IV 41; BGE 97 II 97.

[48] GE 138 II 346 E. 8.2; BGE 70 II 130; BGE 127 III 492; BGE 138 II 346 E. 8.3.

[49] Siehe dazu allgemein die Kapitel 6 und 7.

letzung dar.[50] Die Veränderung von bestehenden Aufnahmen mittels technischer Hilfsmittel berührt ebenfalls das Recht am eigenen Bild und häufig auch den Schutz der Ehre (Nobel/Weber 2021: 312). Das Recht an der eigenen Stimme wird durch die unautorisierte Verwendung in Deepfakes verletzt (ebd.: 312). Zudem wird durch die Anmassung der Individualität im Deepfake der Namensschutz verletzt (ebd.: 314). Werden Deepfakes verwendet, um verstorbene Personen darzustellen, können sich die Nachkommen bis zu einem gewissen Grad auf den eigenen Persönlichkeitsschutz berufen, um solche Darstellungen zu verhindern (Cappello 2020: 19 ff.; Breitschmid/Kamp).

Im Ergebnis kann also gesagt werden, dass das Erstellen eines Deepfakes einer konkreten Person *in jedem Fall* eine Persönlichkeitsverletzung darstellt. Fraglich ist nur, ob ein *Rechtfertigungsgrund* vorgebracht werden kann.

Zulässige Rechtfertigungsgründe einer Persönlichkeitsverletzung sind die Einwilligung der betroffenen Person, überwiegende private oder öffentliche Interessen sowie Gesetzesvorschriften.[51] Kann kein Rechtfertigungsgrund geltend gemacht werden, liegt eine widerrechtliche Persönlichkeitsverletzung vor. Ein Verschuldensnachweis ist nicht notwendig (Nobel/Weber 2021: 283).

- Ein denkbarer Rechtfertigungsgrund ist die *Einwilligung* des Betroffenen für das Erstellen eines Deepfakes. Dies könnte etwa im Rahmen eines Kunstprojekts geschehen.
- Denkbar wäre auch eine *Gesetzesvorschrift*, die das Erstellen von Deepfakes zum Zwecke der Strafverfolgung, etwa zur Rekonstruktion von Tathergängen, erlaubt.[52] Eine solche existiert derzeit aber in der Schweiz nicht.
- Je nach Umständen wäre auch ein *überwiegendes öffentliches Interesse* zu prüfen. Denkbar ist eine Rechtfertigung über die Kunst-, Meinungs- und Informationsfreiheit unter anderem für Satire (Kuhlmann 2020). Diesen Grundrechten kommt öffentliches Interesse zu (Nobel/Weber 2021: 328).[53] Das Bundesgericht prüft das Vorliegen von Satire anhand der drei Elemente Aggression, Soziales und Ästhetisches. Das aggressive Element sollte sich nicht gegen eine bestimmte Person richten, sondern dem Repräsentanten

[50] BGE 129 III 715 E. 4.1

[51] Art. 28 Abs. 2 ZGB.

[52] Dazu auch unten, 4.3 (Deepfakes vor Gericht).

[53] BGer Urteil 5A.553/2013 vom 14. April 2014 E. 3.2.

eines bestimmten Verhaltens, einer Ordnung oder einer Institution gelten. Das soziale Element beinhaltet beispielsweise die Aufdeckung eines Widerspruchs. Das ästhetische Element erfordert die Verwendung von Stilmitteln (ebd.: 328; Cueni).[54] Letztlich ist für die Qualifikation eine Interessenabwägung im Einzelfall vorzunehmen.[55] Für Deepfakes bedeutet dies: Ein klar als Satire erkennbarer Deepfake, der nicht wie eine echte Darstellung daherkommt, wird eher zu rechtfertigen sein als der Fall einer täuschend echten Darstellung, in welcher das Interesse des Betroffenen am Persönlichkeitsschutz überwiegen würde.

Abschliessend lässt sich festhalten, dass mit wenigen Ausnahmen (Einwilligung der Betroffenen, Gesetzesvorschrift, Satire) die durch ein Deepfake begangene Persönlichkeitsverletzung nicht zu rechtfertigen und somit rechtswidrig ist.

#### 4.2.1.2. Datenschutz

Die Bestimmungen des Datenschutzes ergänzen und konkretisieren den Persönlichkeitsschutz. Personendaten sind alle Angaben, die sich auf eine bestimmte oder bestimmbare natürliche Person beziehen. Ist das Gesicht einer Person in einem Deepfake erkennbar oder ist die Person anderweitig bestimmbar, liegen damit Personendaten vor und das Datenschutzgesetz findet Anwendung (siehe zur Anwendbarkeit der DSGVO: van Huijstee u.a. 2021: 38 f.).

Wie aufgezeigt, stellt sowohl das Erstellen wie auch das Verbreiten von Deepfakes ohne die Einwilligung der Betroffenen meist eine *widerrechtliche Persönlichkeitsverletzung* dar, dies trifft auch auf den datenschutzrechtlichen Begriff der Persönlichkeitsverletzung (Art. 30 Abs. 2 DSG) zu. Zudem würden dadurch verschiedene Datenschutzgrundsätze (Rechtmässigkeit, Treu und Glauben, Zweckbindung, Erkennbarkeit der Datenbeschaffung und Datenrichtigkeit) oder auch anderweitige Vorschriften, etwa betreffend die Bekanntgabe an Dritte von besonders schützenswerten Personendaten (bspw. Darstellungen über die Intimsphäre oder politische Äusserungen), verletzt.[56]

---

[54] BGer Urteil 5A.553/2013 vom 14. April 2014 E. 3.3.

[55] BGE 70 II 130; BGE 127 III 492; BGer Urteil 5A_376/2013 vom 29. Oktober 2013 E. 4.3 f.; BGE 138 III 641 E. 4.4.2; EGMR-Urteil vom 9. Januar 2018, GRA Stiftung gegen Rassismus und Antisemitismus gegen die Schweiz, Rz. 65; BGer 5A_553/2012 vom 14. April 2014 E. 3.5 f.

[56] Art. 30 Abs. 2 lit. c DSG.

Personendaten werden nicht nur bei der Erstellung und Verbreitung eines spezifischen Deepfakes verwendet, sondern werden meist bereits im Vorfeld für die Erstellung der Software benötigt (van Huijstee u.a. 2021: 39). Auch hier stellt sich die Frage, inwiefern die verwendeten Daten rechtmässig erhoben werden. Wird eine Deepfake-Software auf dem Markt angeboten, hat der Verantwortliche für die Datenverarbeitung im Voraus eine Datenschutzfolgeabschätzung durchzuführen (ebd.: 39).

Die Verletzung der Datenschutzvorschriften kann, wie oben bezüglich Persönlichkeitsverletzungen beschrieben, aufgrund des Vorbringens von *Rechtfertigungsgründen* zulässig sein.[57] Die Einwilligung des Betroffenen stellt einen zulässigen Rechtfertigungsgrund dar. Da von beiden Daten bearbeitet werden, müsste einerseits die Einwilligung der Person, deren Daten original verwendet werden, und anderseits der Person, deren Daten im neuen Medium verwendet werden, vorliegen (van Huijstee u.a. 2021: 39; Hewage 2020). Ein überwiegendes Interesse an der Verbreitung eines Deepfakes könnte beispielsweise bei der satirischen Darstellung einer berühmten Person vorgebracht werden. Entsprechend wäre das Vorliegen eines Rechtfertigungsgrunds im Einzelfall abzuklären. In der Regel wird jedoch gegen eine Vielzahl von datenschutzrechtlichen Vorschriften verstossen, ohne dass ein zulässiger Rechtfertigungsgrund ersichtlich wäre.

Aus dem Datenschutzrecht ergeben sich zudem *Sorgfaltspflichten*, welche Verantwortliche, die Personendaten bearbeiten, einzuhalten haben. Gemäss dem Grundsatz der Datensicherheit sind angemessene organisatorische und technische Massnahmen zu treffen, um den unautorisierten Zugang zu Personendaten zu verhindern. Aufgrund der Möglichkeiten, die sich aufgrund der Deepfake-Technologie ergeben, müssen auch allfällige Identifikations- oder Authentifikationsverfahren mittels Face-ID oder Spracherkennung auf ihre Sicherheit überdacht werden. Organisatorische Massnahmen wie Schulungen des Personals erscheinen angezeigt, um Betrugsversuchen zuvorzukommen (vgl. auch die unter 7.4 diskutierten Massnahmen).

#### 4.2.1.3. Urheberrecht

Bei der Erstellung von Deepfakes wird in der Regel bereits bestehendes Bild- oder Tonmaterial als «Input» verwendet. Die Verwendung von fremdem Bild-

[57] Art. 31 DSG.

und/oder auch Tonmaterial ohne Zustimmung der Urheberin stellt eine Rechtsverletzung dar, denn diese hat das ausschliessliche Recht zu entscheiden, ob ihr Werk geändert oder zur Schaffung eines neuen Werks verwendet werden darf.[58] Auch bei einem vertraglich eingeräumten Recht kann sich die Urheberin gegen die Weiterverwendung wehren, insbesondere im Falle einer Persönlichkeitsverletzung. Eine Ausnahme besteht für die Schaffung von Parodien und vergleichbaren Werken.[59] Jedoch sind auch hier die Schranken des Persönlichkeitsschutzes zu beachten.

Das Verwenden von Werken für den Privatgebrauch ist ebenfalls zulässig.[60] So kann die Verwendung eines Werks im Internet – unter der Voraussetzung, dass es sich nur um eine sehr kleine, von der Öffentlichkeit abgegrenzte Gruppe handelt – rechtmässig sein,[61] wobei es in der Praxis zu schwierigen Abgrenzungsfragen kommen dürfte, wann eine Gruppe (z.B. ein geschlossenes Onlineforum oder eine Telegram-Gruppe) noch als von der Öffentlichkeit abgegrenzt betrachtet werden kann.

Bezogen auf Deepfakes bedeutet das, dass die Erstellung und Verwendung von Deepfakes urheberrechtlich zulässig ist, sofern der Deepfake nicht über einen kleinen Personenkreis hinausgehend verbreitet wird oder wenn es sich um eine Parodie handelt. In allen anderen Fällen dürfte eine Urheberrechtsverletzung vorliegen.

### 4.2.2. Schutz im Strafrecht

Wie gesagt, ist das Erstellen oder Verbreiten von Deepfakes in der Schweiz nicht *per se* strafbar (Jacquemin 2023: 325). Gewisse Formen von Deepfakes bzw. deren Verwendung stellen jedoch nicht nur eine (zivilrechtliche) Persönlichkeits-, Datenschutz- oder Urheberrechtsverletzung dar, sondern eine strafbare Handlung. Deepfakes können etwa verwendet werden, um dem Ansehen einer Person zu schaden, diese zu mobben, einzuschüchtern, zu erpressen oder ihre Identität zu stehlen. Auch die Verwendung von Deepfakes zur Kursmanipulation an der Börse ist strafbar.

---

[58] Art. 11 URG.

[59] Art. 11 Abs. 3 URG.

[60] Art. 19 Abs. 1 lit. a URG.

[61] Art. 10 Abs. 2 lit. a und c URG.

#### 4.2.2.1. Anwendbarkeit des Schweizer Strafrechts

Für die Anwendbarkeit des schweizerischen Strafrechts ist ein ausreichender Bezug zur Schweiz nötig. Das kann bei Delikten, die über das Internet getätigt werden, Schwierigkeiten bereiten. Ein möglicher Anknüpfungspunkt für die Strafbarkeit einer Handlung ist der Ort, an welchem das Delikt begangen wurde.[62] Das Delikt gilt als dort begangen, wo der Erfolg eintritt.[63] Handelt es sich um Ehrverletzungen, ist das schweizerische Strafrecht anwendbar, wenn sie in der Schweiz geäussert werden. Befindet sich die Täterschaft im Ausland und verletzt die Ehre einer in der Schweiz wohnhaften Person, ist die Anwendbarkeit des StGB fraglich, da Ehrverletzungsdelikte als abstrakte Gefährdungsdelikte konstruiert sind. Der Begriff des Erfolgs wird in der Rechtslehre und -praxis jedoch weit ausgelegt, weshalb bei Ehrverletzungsdelikten die Anwendbarkeit des StGB in der Tendenz dennoch bejaht wird (Robin/Christof 2022: 1249–1259).

#### 4.2.2.2. Identitätsmissbrauch

Im Zuge der Totalrevision des Schweizerischen Datenschutzgesetzes wurde das Strafrecht um den Straftatbestand des Identitätsmissbrauches ergänzt. Der neue Art. 179$^{decies}$ StGB soll die Persönlichkeit des Individuums schützen; das Recht auf Respektierung und Achtung der Identität wird damit strafrechtlich verankert. Der Täter erfüllt den Straftatbestand, wenn er die Identität mit einer Absicht einer Schädigung oder Vorteilsbeschaffung verwendet. Vorausgesetzt wird, dass dem Betroffenen durch den Identitätsmissbrauch ein Nachteil entstanden ist. Dieser muss eine gewisse Schwere aufweisen und kann materieller oder immaterieller Natur sein. Bereits das massive Ärgern eines Betroffenen kann das Erfordernis der Nachteilsabsicht erfüllen.[64] Mit welchen Mitteln ein Identitätsmissbrauch begangen wurde, ist unerheblich. Eine Handlung aus reinem Übermut ist nicht strafbar.

Diese neue Strafnorm steht u.a. unter Kritik, weil bereits viele andere Tatbestände die strafbaren Handlungen erfassen und sich demnach Abgrenzungsschwierigkeiten ergeben werden (Reber 2020: 33 ff.). Wird beispielsweise die Identität einer Person mit Schädigungsabsicht oder zur Erlangung eines unrechtmäs-

---

62 Art. 3 Abs. 1 StGB.

63 Art. 8 StGB.

64 BBl 2017 6941, 7127 f.

sigen Vorteils verwendet, ist die Anwendung anderer Strafbestimmungen wie Betrug, Urkundenfälschung oder Ehrverletzung zu prüfen.

#### 4.2.2.3. Ehrverletzungsdelikte

Abgesehen von zivilrechtlich relevanten Persönlichkeitsverletzungen (oben, Kapitel 4.2.1.1) können Deepfakes auch eine strafrechtlich relevante Ehrverletzung gemäss Art. 173 ff. StGB darstellen.

Die ehrverletzende Äusserung kann sowohl gegenüber der betroffenen Person als auch Dritten erfolgen. In ersterem Fall ist der Tatbestand der Beschimpfung (Art. 177 StGB) zu prüfen, in letzterem Fall kann es sich um üble Nachrede oder Verleumdung (Art. 173 oder 174 StGB) handeln. Die verwendeten Mittel sind unerheblich (Riklin 2019b: N 1).[65] Eine Ehrverletzung wird nicht von Amtes wegen verfolgt; Betroffene müssen selber Strafanzeige erstatten.

Fraglich ist, wann das Rechtsgut der Ehre verletzt ist. Nach Lehre und Rechtsprechung geht es um den Ruf und die Wertschätzung einer Person als ehrbarer Mensch und ihre Geltung bei Dritten. Damit ist insbesondere die Bezichtigung moralisch verwerflicher Handlungen strafbar.[66] Ehrverletzungen können als Tatsachenbehauptungen, Werturteile oder gemischte Werturteile geäussert werden (Riklin 2019a).

Schwierig ist die Abgrenzung zwischen ehrverletzenden und «lediglich» peinlichen oder verfälschten Darstellungen, insbesondere bei *deepnudes* oder *deepfake porn*. Eine bildliche Darstellung aus dem Intimbereich der abgebildeten Person ist gemäss Bundesgericht für sich allein nicht ehrverletzend.[67] Ist die Aufnahme aber ehrenrührig oder so verfälscht, dass aus den Umständen eine Verunglimpfung oder Blossstellung der betroffenen Person hervorgeht, können die Ehrverletzungsdelikte erfüllt sein (Bundesrat 2021: 21). Ist die Aufnahme dagegen lediglich freizügig, lässt sich deren Weiterleitung nicht unter die Ehrverletzungsdelikte subsumieren (ebd.: 23). Eine Ehrverletzung dürfte zu bejahen sein, wenn der Deepfake suggeriert, die betroffene Person betätige sich als Pornodarsteller, Prostituierte oder Zuhälter; auch die Unterstellung, man unterhalte eine Aussenbeziehung, ist i.d.R. ehrverletzend (Tag/Wyss 2024). Es stellt sich dabei aber das

---

[65] Art. 176 StGB.

[66] BGE 76 IV 29.

[67] SJZ 2004, 95 f.

Problem, dass Deepfakes mit sexualisierten Inhalten nicht alleine die Ehre verletzen, sondern auch die sexuelle Selbstbestimmung, was durch eine reine Subsumtion unter die Ehrverletzungsdelikte nicht berücksichtigt wird. Allenfalls kann aber nunmehr der neue Straftatbestand des unbefugten Weiterleitens von nicht öffentlichen sexuellen Inhalten (Art. 197a StGB, s.u. 4.2.2.5) einschlägig sein.

In Bezug auf Deepfakes ist des Weiteren der Straftatbestand der *üblen Nachrede* (Art. 173 StGB) zu prüfen. Dieser ist einerseits durch den Vorwurf unehrenhaften Verhaltens oder anderseits durch Äusserung rufschädigender Tatsachen jeweils gegenüber Dritten erfüllt.[68] Auch Weiterverbreitung von Rufschädigungen sind unzulässig. Massgebend für das Vorliegen der Rufschädigung ist der Eindruck eines Durchschnittspublikums. Der Täter kann sich mittels Nachweises des guten Glaubens oder Wahrheitsbeweises vom Vorwurf einer üblen Nachrede entlasten. Bezweckt die Äusserung nicht die Verfolgung eines öffentlichen Interesses und wird vor allem mit der Absicht gestreut, jemandem etwas Übles vorzuwerfen, ist das Vorbringen eines Entlastungsbeweises nicht möglich (siehe hierzu: Riklin 2019a). Handelt der Täter wider besseres Wissen, ist der Tatbestand der Verleumdung erfüllt.[69]

Eine üble Nachrede oder Verleumdung kann auch begangen werden, wenn jemand einen rufschädigenden Deepfake weiterverbreitet (Art. 173 und 174 Ziff.1 Abs. 2 StGB). Gemäss Bundesgericht macht sich jemand, der einen diffamierenden Beitrag auf Facebook *likt* oder teilt, unter Art. 173 Ziff. 1 Abs. 2 StGB strafbar.[70] Gleiches dürfte auf rufschädigende Deepfakes zutreffen. Dabei wird unterschieden, ob die weiterverbreitende Person wusste, dass es sich um eine Fälschung handelt (womit der Tatbestand der Verleumdung zu prüfen wäre) oder nicht (was einer üblen Nachrede gleichkommt) (Jacquemin 2023: 233).

#### 4.2.2.4. Drohung, Nötigung, Erpressung, Betrug

Auch das Erfüllen des Tatbestandes einer Drohung (Art. 180 StGB) oder Nötigung (Art. 181 StGB) mittels eines Deepfakes wäre abhängig vom beabsichtigten Ziel erfüllt.[71] Es liegt eine *Drohung* vor, wenn dem Opfer ein künftiges Übel

---

[68] Art. 173 StGB.

[69] Art. 174 StGB.

[70] BGE 146 IV 23.

[71] Art. 180/181 StGB.

angekündigt oder in Aussicht gestellt und dies den Betroffenen ängstigt. Es geht um schwerwiegende Angriffe auf das innere Gleichgewicht einer Person (Delnon/Rüdy 2019: 12 ff.). Die Drohung kann auch Rechtsgüter Dritter betreffen.

*Nötigung* bedeutet die Gewaltanwendung oder Androhung ernstlicher Nachteile, um die Freiheit der Willensbildung und damit die Entscheidungsintensität einer Person einzuschränken. Denkbar wäre auch eine *Erpressung* (Art. 156 StGB), was eine qualifizierte Form der Nötigung ist.[72]

Deepfakes könnten also zur Drohung, Nötigung oder Erpressung einer Person verwendet werden. Je nach dargestelltem Inhalt in Deepfake-Videos könnten auch weitere Strafnormen verletzt sein. Denkbar ist die Nötigung von Drittpersonen mittels Deepfakes, um diese etwa zu einem Geldtransfer, zur Bekanntgabe von Geschäftsgeheimnissen oder Passwörtern zu bringen (vgl. auch die Kapitel 6 und 7).

Die arglistige Irreführung einer Person mit der Absicht, sich oder andere unrechtmässig zu bereichern, erfüllt den Straftatbestand des *Betrugs* (Art. 146 StGB).[73] Wird ein Deepfake verwendet, um jemanden zu imitieren und damit einer Person vertrauliche Informationen zu entlocken oder sie zu einer Geldüberweisung zu animieren («Telefontrick»), könnte ein Betrug vorliegen. Die für einen Betrug nötige Arglist liegt vor, wenn ein Täter absichtlich technische Manipulationen vornimmt, um eine Drittperson in die Irre zu führen.[74] Keine Arglist liegt vor, wenn der Irrtum mit einem Mindestmass an Aufmerksamkeit hätte vermieden werden können. Dies ist im Einzelfall anhand der vorliegenden Umstände zu bewerten. Es geht hier also um die Qualität der Täuschung (Mäder/Niggli 2019: N 69). Ein sehr fehlerhaftes, offensichtliches Deepfake würde allenfalls nicht für die Arglist ausreichen. Hier müsste aber insbesondere auch die Erfahrenheit des Opfers berücksichtigt werden. Auch ein vorgespielter Zeitdruck oder eine fiktive Notlage müssten beachtet werden.[75] Betrug ist ein Delikt, welches von Amtes wegen verfolgt wird.

Der betrügerische *Missbrauch einer Datenverarbeitungsanlage* zur unrechtmässigen Bereicherung ist ebenfalls strafbar.[76] Denkbar ist dies etwa, wenn die

---

[72] Art. 156 StGB.

[73] Art. 146 StGB.

[74] Siehe dazu BGE 142 IV 153, 154 f.; BGE 135 IV 76, 81 f.

[75] Siehe dazu BGE 142 IV 153, 135 IV 76, 80 f.

[76] Art. 147 StGB.

Identität einer Person überzeugend gefälscht wurde, um einen biometrischen Identifizierungsprozess, z.B. beim Onlinebanking, zu umgehen. Je nach Absicht des Täters kommt auch eine arglistige Vermögensschädigung infrage.[77]

#### 4.2.2.5. Pornografie oder sexuelle Belästigung

Da ein überwiegender Grossteil von Deepfakes pornografische Inhalte darstellt, ist auch ein Verstoss gegen den Straftatbestand der *Pornografie* (Art. 197 StGB) zu prüfen.[78] Gemäss Art. 197 Abs. 2 StGB wird bestraft, wer pornografische Schriften, Ton- oder Bildaufnahmen oder Abbildungen öffentlich ausstellt oder zeigt oder sie sonst jemandem unaufgefordert anbietet, wobei sich die Strafbarkeit unterscheidet, je nachdem ob der Empfängerkreis nur Erwachsene oder auch Minderjährige umfasst. Dabei spielt es keine Rolle, ob es sich um manipulierte «echte» oder um gänzlich synthetisch generierte Aufnahmen handelt. Für das Erfüllen des Tatbestands ist es notwendig, dass die Pornografie öffentlich ausgestellt oder unaufgefordert angeboten wird. Dies dürfte etwa der Fall sein, wenn Internetuser unerwartet mit einem Deepfake mit pornografischem Inhalt konfrontiert werden, z.B. wenn ein solcher Inhalt beim Durchscrollen des Facebook- oder Instagram-Newsfeeds erscheint (Jacquemin 2023: 327). Hingegen ist der Tatbestand nicht erfüllt, wenn jemand etwa ein *deepnude* ausschliesslich an das Opfer schickt oder wenn ein mittels Deepfakes erstellter pornografischer Inhalt auf einem geschlossenen Forum gezeigt wird (ebd.: 326 f.).

In diesen Fällen kann allenfalls der Tatbestand der sexuellen Belästigung (Art. 198 StGB) erfüllt sein. Im Rahmen der Revision des Sexualstrafrechts soll der Tatbestand neu auch durch die Begriffe «Schrift» und «Bild» ergänzt werden. Das Bundesgericht hat in einem neueren Urteil[79] festgehalten, Artikel 198 Absatz 2 StGB umfasse nicht nur ausgesprochene Worte, sondern auch schriftliche oder bildliche Tatobjekte. Damit wird namentlich das Versenden sexuell konnotierter Bilder oder anzüglicher Nachrichten vom Tatbestand erfasst.[80] Somit dürfte auch das Versenden eines pornografischen Deepfakes entweder an die darauf abgebildete Person oder an eine andere Person eine sexuelle Be-

---

[77] Art. 151 StGB.

[78] Art. 197 StGB.

[79] Urteil des Bundesgerichts 6B_69/2019 vom 4. November 2019, E. 2.3.2.

[80] BBl 2022 687, 60 f.

lästigung darstellen, sofern diese Belästigung durch Schrift oder Bild in grober Weise erfolgt (Bundesrat 2021: 18; Tag/Wyss 2024, Rz. 35 f.

Als Auffangtatbestand für Belästigungen, die nicht unter die genannten Tatbestände fallen, kann der Missbrauch einer Fernmeldeanlage dienen. Bestraft wird, wer aus Bosheit oder Mutwillen eine Fernmeldeanlage zur Beunruhigung oder Belästigung missbraucht. Auch obszöne Belästigungen fallen unter den Tatbestand. Damit ist dieser geradezu auf Belästigungen im Internet zugeschnitten (Bundesrat 2021: 19). Massgebend ist, dass die Kontaktaufnahmen mittels elektronischer Kommunikationsmittel in objektiver Hinsicht lästig oder beunruhigend sind und eine gewisse minimale, quantitative Intensität und/oder qualitative Schwere erreichen.[81]

Im Rahmen der Revision des Sexualstrafrechts wurde zudem ein neuer Art. 197a StGB («Unbefugtes Weiterleiten von nicht öffentlichen sexuellen Inhalten») aufgenommen, hauptsächlich um das Phänomen der Rachepornografie in einem Spezialtatbestand zu kriminalisieren. Bestraft wird nunmehr, wer einen nicht öffentlichen sexuellen Inhalt ohne Zustimmung der darin erkennbaren Person weiterleitet. Sofern es sich um eine erkennbare Person handelt, ist unseres Erachtens dieser Tatbestand auch auf mittels Deepfakes manipulierte Aufnahmen anwendbar. Somit dürfte auch die Konstellation erfasst sein, wo ein pornografischer Deepfake in einem geschlossenen Forum geteilt wird (und somit der Tatbestand der Pornografie gem. Art. 197 StGB nicht greift), dies jedenfalls mindestens, insoweit der Deepfake so «gut» gemacht ist, dass er geeignet ist, das Publikum zu täuschen.

#### 4.2.2.6. Kinderpornografie

Das Herstellen, Besitzen, Verbreiten oder Vorführen von Aufnahmen, die sexuelle Handlungen mit Kindern darstellen, ist in Art. 197 Abs. 4 und 5 StGB unter Strafe gestellt. Pornografische Deepfakes von Kindern, seien sie durch «Ausziehen» *(deepnudes)* einer Darstellung eines Kindes oder durch Verwenden des Gesichts eines Minderjährigen in einer pornografischen Darstellung, fallen unter die verschärften Strafbestimmungen von Art. 197 Abs. 4 und 5 StGB (Jacquemin 2023: 328 f.), unabhängig davon, ob die Darstellungen von «echten» Minderjährigen stammen oder synthetisch generiert wurden.

[81] BGE 126 IV 216 E. 2.

Noch ungeklärt in Lehre und Rechtsprechung ist die Frage, ob etwa beim Einfügen eines Gesichts auf eine vorhandene Aufnahme oder auch bei vollständig synthetisch generierten Inhalten die Tatbestandsvariante der «nicht tatsächlichen sexuellen Handlungen mit Minderjährigen» erfüllt ist, welche einen geringeren Strafrahmen aufweist, als wenn die Aufnahmen tatsächliche sexuelle Handlungen mit Minderjährigen zeigen.

Würden Deepfakes zwecks *Cyber-Grooming*, also zur gezielten Kontaktaufnahme Erwachsener mit Minderjährigen in Missbrauchsabsicht,[82] verwendet, beispielsweise indem ein Erwachsener sich auf Bildern oder Videoaufnahmen als Jugendlicher darstellen lässt, so kann ein strafbarer Versuch, sexuelle Handlungen mit Kindern vorzunehmen (Art. 187 Ziff. 1 erster Absatz i.V.m. Art. 22 StGB) oder Kinderpornografie herzustellen (Art. 197 Abs. 4 zweiter Satz i.V.m. Art. 22 StGB), vorliegen, sofern es zu einem Treffen kommt und dieses den letzten entscheidenden Schritt auf dem Weg zur Tatverwirklichung darstellt.[83]

#### 4.2.2.7. Gezielte Desinformation der Öffentlichkeit

Ein Deepfake-Video könnte auch das Sicherheitsgefühl der Bevölkerung als Rechtsgut verletzen. So wird gemäss Strafgesetzbuch eine Person bestraft, welche die Bevölkerung durch Androhung oder Vorspielen einer Gefahr für Leib, Leben oder Eigentum in Schrecken versetzt (Art. 258 StGB). Die gezielte Desinformation der Bevölkerung ist in der Regel nicht strafbar, jedoch gibt es einige Ausnahmen, insbesondere wenn eine Desinformation sehr glaubwürdig erscheint. Für diesen Straftatbestand wird vorausgesetzt, dass der Täter etwas bewusst vorspielt, also beispielsweise eine Gefahr simuliert (Fiolka 2019: N 15). Auch kann die öffentliche Aufforderung zu Verbrechen oder Vergehen mit Gewalttätigkeit gegen Menschen oder Sachen bestraft werden.[84]

Ein Deepfake, das darauf gerichtet ist, die verfassungsmässige Ordnung der Eidgenossenschaft oder der Kantone rechtswidrig zu stören oder zu ändern, würde den Straftatbestand gemäss Art. 275 StGB erfüllen. Auch wer eine Versammlung, Wahl oder Abstimmung durch Androhung ernstlicher Nachteile hin-

---

[82] Vgl. Kommission für Rechtsfragen des Ständerates, Strafrahmenharmonisierung und Anpassung des Nebenstrafrechts an das neue Sanktionenrecht. Vorlage 3: Bundesgesetz über eine Revision des Sexualstrafrechts, BBl 2022 687, 70.

[83] BBl 2022 687, 71.

[84] Art. 259 StGB.

dert oder stört, kann bestraft werden.[85] Denkbar wäre auch die Beeinträchtigung internationaler Beziehungen mittels Anwendungen von Deepfake; hier wäre der Straftatbestand der Beleidigung eines fremden Staates oder zwischenstaatlicher Organisationen zu prüfen.[86]

Werden gefälschte Videos als Deepfakes in gerichtlichen Verfahren vorgebracht, ist dies ebenfalls strafbar (siehe Delnon/Rüdy 2019: 5).[87] Die Ermittlung der Echtheit eines Beweises könnte jedoch zu Problemen führen (siehe dazu auch unten, Kapitel 4.3.1). Angeschuldigte müssen gemäss Art. 6 EMRK ausreichende Möglichkeiten haben, vorzubringen, dass Beweise allfällig gefälscht sind.[88] Es obliegt jedoch den Mitgliedstaaten, über Regeln der Beweiszulassung zu entscheiden. Die Modalitäten in der Schweiz werden im Kapitel zur Rechtsdurchsetzung besprochen.

Denkbar wäre auch, dass beispielsweise falsche Informationen zu Übernahme oder Zusammenschluss von Unternehmen oder Verlustgeschäft publiziert würden. Der Kurs von börsenkotierten Unternehmen könnte dadurch unzulässig manipuliert werden (van Huijstee u.a. 2021: 31). Die Verbreitung von falschen oder irreführenden Informationen wider besseres Wissen zur Kursmanipulation ist nach Art. 155 FinfraG strafbar.[89]

#### 4.2.2.8. Strafbare Vorbereitungshandlungen

Vorbereitende Handlungen im Zusammenhang mit dem Erstellen von Deepfakes können strafbar sein, wenn eine unrechtmässige Beschaffung von fremden Daten und damit eine unbefugte Datenbeschaffung nach Art. 143 StGB vorliegt. Ebenfalls strafbar ist das Eindringen in ein Computersystem für das Erlangen von persönlichen Daten. Werden zur Erstellung des Deepfakes heimlich Fotos der betroffenen Person gemacht, ist auch hier die Strafbarkeit der Handlung möglich. Auch strafbar ist das Beschaffen von besonders schützenswerten und nicht frei zugänglichen Personendaten (Ramel/Vogelsang 2019: 21 ff.).

---

[85] Art. 279 StGB.

[86] Art. 296 und Art. 297 StGB.

[87] Art. 307 StGB.

[88] EGMR-Urteil vom 06.06.2002 «Karalevičius v Lithuania».

[89] Bundesgesetz über die Finanzmarktinfrastrukturen und das Marktverhalten im Effekten- und Derivatehandel (FinfraG), SR 958,1.

### 4.2.3. Verfahrensrechtliche Geltendmachung

Neben der Frage, ob und welche Rechtsansprüche durch einen Deepfake verletzt sind, ist auch zu untersuchen, wie Betroffene dagegen vorgehen können. Zu unterscheiden sind drei Konstellationen (Rosenthal 2014: 415 f.), abhängig von folgenden zwei Fragen: Ist der Urheber der Persönlichkeitsverletzung bekannt oder ist er unbekannt? Liegt ein strafbares Verhalten vor oder «nur» eine zivilrechtliche Persönlichkeitsverletzung?

#### 4.2.3.1. Urheber bekannt

Zivilrechtliche Möglichkeiten bei Verletzung der Persönlichkeit sind die Unterlassungsklage, die Beseitigungsklage sowie die Feststellungsklage.[90] Vor Einreichung einer Persönlichkeitsschutzklage ist es empfehlenswert, die verantwortliche Person zur Löschung innert einer gegebenen Frist aufzufordern. Bleibt die Reaktion aus, kann die Löschung allenfalls über den Betreiber der Plattform erreicht werden (Rosenthal 2014: 416).

Das Unterlassungsbegehren richtet sich gegen den Verursacher der Persönlichkeitsverletzung; dieser soll angewiesen werden, die Verletzung (auch in Zukunft) zu unterlassen. Mit dem Beseitigungsbegehren kann die Löschung eines bestimmten, konkret bezeichneten, rechtswidrigen Inhalts durchgesetzt werden. Dabei kann sowohl gegen den Urheber des Deepfakes als auch gegen die Plattform, die ihn beherbergt, vorgegangen werden (Jacquemin 2023: 329).[91] Im Falle eines Feststellungsbegehrens wird das Vorliegen eines rechtswidrigen Inhalts vom Gericht bestätigt. Dies kann für eine Rechtsdurchsetzung im Ausland nötig sein (Rosenthal 2014: 419).

Neben der Löschung bzw. der Feststellung der Rechtswidrigkeit können ggf. pekuniäre Ansprüche geltend gemacht werden. Der durch die Persönlichkeitsverletzung verursachte seelische Schmerz kann in Form einer Genugtuung beim Verursacher der Rechtsverletzung eingeklagt werden. Resultiert die Rechtsverletzung in einer Vermögenseinbusse, kann Schadenersatz verlangt werden. Schwieriger ist es, wenn durch den Deepfake ein Reputationsverlust eingetreten ist. Sofern dieser sich nicht zahlenmässig beziffern lässt, gilt er nicht als Schaden im Rechtssinne (Rosenthal 2014: 419).

---

[90] Art. 28a ZGB.

[91] Mit Verweis auf Urteil des BGer 5A_792/2011 vom 14. Januar 2011 E. 6.3.

Gerichtsverfahren können langwierig sein. Um die Verbreitung des infrage stehenden Deepfakes während des Gerichtsverfahrens zu unterbinden, ist allenfalls die Beantragung von *vorsorglichen Massnahmen* zu prüfen. Sieht das Gericht nach Anhörung der beiden Parteien gute Erfolgsaussichten, gewährt es vorsorgliche Massnahmen. Die Frage, ob die Folgen für den Beklagten durch die Anordnung für die Dauer des Verfahrens vertretbar wären oder ob ein allfälliger nicht wiedergutzumachender Nachteil entstünde, wird berücksichtigt. Das hierfür nötige Verfahren dauert einige Wochen. Ohne vorherige Anhörung der Gegenpartei und damit umgehend umsetzbar sind *superprovisorische Massnahmen*. Diese können etwa bei unmittelbar bevorstehenden Persönlichkeitsverletzungen verlangt werden (Rosenthal 2014: 420). Da Suchmaschinen Zwischenspeicher betreiben, kann es angezeigt sein, die Entfernung eines Treffers von der Suchmaschine zu verlangen. Viele Suchmaschinen bieten an, eine bestimmte Seite auch vorzeitig zu aktualisieren und so Treffer aus den Suchresultaten zu entfernen (ebd.: 421).

Das grenzüberschreitende Durchsetzen von vorsorglichen Massnahmen gestaltet sich häufig kompliziert. Superprovisorische Massnahmen sind vom Lugano-Übereinkommen (LugÜ) ausgenommen, weshalb das erleichterte Anerkennungs- und Vollstreckungsverfahren nicht zur Anwendung gelangt.

Im Zivilverfahren hat die klagende Partei die Verfahrenskosten im Voraus zu zahlen. Im Anschluss an das Gerichtserfahren werden sie der unterliegenden Partei auferlegt.

Führt die Verbreitung eines Deepfakes zu einer *Urheberrechtsverletzung*, kann die Urheberin in einem zivilrechtlichen Verfahren mittels Leistungsklage die Beseitigung des Deepfakes verlangen.[92] Weiter wären die Feststellungsklage sowie Klagen auf Schadenersatz oder Genugtuung zu prüfen (Heinemann/Althaus 2015: 7). In der Praxis dürften urheberrechtliche Rechtsansprüche jedoch kaum durchsetzbar sein, da oft unzählige Fotografien von verschiedenen Urhebern verwendet werden. Verfahrenstaktisch sind deshalb Klagen des Persönlichkeitsschutzes oder Datenschutzrechtes naheliegender, jedenfalls wenn – wie im Grossteil der im Internet kursierenden Deepfakes der Fall – reale Personen abgebildet sind. Dies ist allerdings unbefriedigend, wenn der Urheberschutz am Werk (was häufig der Fall sein dürfte) einer anderen als der abgebildeten Person (z.B. dem Fotografen) zukommt (Cappello 2020: 162).

[92] Art. 62 Abs. 1 lit a und b URG.

#### 4.2.3.2. Urheber nicht bekannt

Ist der Urheber der Persönlichkeitsverletzung nicht bekannt, so besteht zivilverfahrensrechtlich keine Möglichkeit, «gegen Unbekannt» zu klagen. In diesem Fall besteht die einzige Möglichkeit für Betroffene darin, sich an den Intermediär resp. die Plattform zu wenden.

Vor der Erhebung von Rechtsmitteln empfiehlt es sich, bestehende *Notice-* und *Take-down*-Verfahren der Plattformen in Anspruch zu nehmen. Betroffene können sich aber auch direkt an Hostingprovider richten. Diese sollten gemäss dem Code of Conduct des Schweizerischen Branchenverbandes Simsa eingegangene Vorwürfe weiterleiten und betreffende Webseiten auffordern, Meldungen zu untersuchen und gegebenenfalls rechtswidrige Inhalte zu entfernen (Swico 2013). Liegt ein eindeutiger Fall vor, kann der Provider den Zugang zur Website vorübergehend sperren (Rosenthal 2014: 416; BAKOM 2011: 64).

Zeigen diese Massnahmen keine Wirkung, kann rechtlich gegen die Betreiber vorgegangen werden. Zur Providerhaftung kennt das Persönlichkeitsrecht jedoch keine ausdrückliche Regelung. Gerichte haben die Verantwortlichkeit von Inhabern privater Webseiten analog zur Verantwortlichkeit von Zeitungen bejaht (Hausheer/Aebi-Müller 2020; BJ 2015; Melli 2022: 55).[93] Dafür reicht bereits die Mitwirkung (auch ein untergeordneter Tatbeitrag) an einer Persönlichkeitsverletzung aus. So wurde ein Zeitungsunternehmen, welches einen Blog betreibt, für das Verhalten eines Bloggers verantwortlich gemacht. Das Bundesgericht befand, eine Blog-Hosterin sei vergleichbar mit einer Zeitung, welche Leserbriefe publiziere. Die Verantwortlichkeit könne unabhängig von der Herrschaft über den Inhalt bestehen (BJ 2015: 32; Rohn 2004: 218).[94] Nicht vorausgesetzt sei die Kenntnis des Providers über das persönlichkeitsverletzende Verhalten. Dieser Entscheid wurde in der Rechtslehre auch kritisiert, insbesondere aufgrund der pauschalen Annahme der Passivlegitimation sowie der im Einzelfall fehlenden Abmahnung der Hosterin (m.w.H. BJ 2015: 32 f.). In einem weiteren Entscheid hat das Bundesgericht klargestellt, dass aus Art. 28 ZGB keine Haftung für fremdes Verhalten abgeleitet werden könne. Eine Mitwirkung durch passives Verhalten setze die Verletzung einer Handelspflicht voraus. Ungenützte Möglichkeiten führten nicht zu einer Verletzung. Das Teilen eines allgemeinen

---

[93] Urteil des Bundesgerichts 5P.308/2003 vom 28. Oktober 2004 E. 2.4.5; BGE 106 II 92; BGE 126 III 161.

[94] BGer vom 14. Januar 2013, 5A_792/2011 E. 6.3.

Website-Links stelle keine Mitwirkung an der Persönlichkeitsverletzung eines nicht explizit verlinkten Berichtes auf der Website dar.[95]

Zusammenfassend ist die Mitwirkung eines Providers und dessen zivilrechtliche Verantwortlichkeit im Einzelfall zu beurteilen. Hierbei hat das Gericht dem Grundsatz der Verhältnismässigkeit Rechnung zu tragen (BJ 2015: 31; Aebi-Müller 2005; Melli 2022: 4). Content-, Access- und Hostingprovidern werden aufgrund unterschiedlicher Inhaltsnähe regelmässig unterschiedliche Verantwortlichkeiten zugesprochen. Eine Passivlegitimation von Hostingprovidern wird grundsätzlich bejaht (BJ 2015: 97).[96] Bei reinen Zugangsvermittlern (Access-Providern) – welche die Leistung automatisiert und weltweit erbringen – wird aufgrund des mangelnden adäquat kausalen Tatbeitrags dies eher zu verneinen sein (ebd.: 98).

Gegen die Plattformbetreiber sind dieselben Klagen wie bei Persönlichkeitsverletzungen durch eine identifizierbare Person einschlägig.[97] Zur Löschung eines Deepfakes kann ein Beseitigungsbegehren gegen den Provider eingereicht werden. Das Vorbringen eines Unterlassungsbegehrens gegen einen Plattformbetreiber ist jedoch umstritten. Die herrschende Lehre verneint die Pflicht zur proaktiven Prüfung der Posts durch die Provider. Provider müssen folgerichtig erst nach der Meldung eines klar rechtsverletzenden Inhalts reagieren. Die Verpflichtung der Plattformen, das Hochladen von illegalen Inhalten generell zu verhindern (Stay-Down-Pflicht), würde einer unzulässigen Vorabkontrolle gleichkommen. Rechtslehre und Rechtsprechung gehen davon aus, dass dies verschiedene Grundrechte, insbesondere den Kerngehalt der Medienfreiheit, betreffen würde (BJ 2015: 45). Eine Verpflichtung von Hostingprovidern zur Verhinderung der Wiederholung einer bestimmten und konkret drohenden Rechtsverletzung wäre denkbar.

In einem älteren Bundesgerichtsentscheid wurde die Mitwirkung einer Druckerei aufgrund des wiederholten Drucks eines ehrverletzenden Artikels bejaht. Die Druckerei hätte nach dem Erscheinen des Artikels die widerrechtliche Persönlichkeitsverletzung bemerken und die weiteren zu druckenden Inhalte besser

---

95 BGer vom 6. Mai 2015, 5A_658/2014 E. 4.2. Siehe auch Rechtsprechung des EGMR: EGMR-Urteil vom 2.2.2016, Magyar Tartalomszolgáltatók Egyesülete/Ungarn, Nr. 22947/13; EGMR-Urteil vom 16.62015, Delfi AS/Estland, Nr. 64569/09.

96 BGer 5A_792/2011 vom 14. Januar 2013.

97 Art. 28a ZGB.

überwachen müssen.[98] Die Übertragbarkeit dieser Entscheidung auf einen Provider bleibt fraglich, da die Dienstleistung weitgehend automatisiert erfolgt und eine riesige Menge an Daten verarbeitet werden (BJ 2015: 46).

Ein Feststellungsbegehren kann allenfalls hilfreich sein, wenn für ein ausländisches Verfahren ein Gerichtsentscheid benötigt wird oder ein kooperationswilliger ausländischer Provider ein entsprechendes Urteil für die Löschung des fraglichen Inhalts verlangt.

Im Unterschied zu den vorgängig genannten Begehren werden für Genugtuung und Schadenersatz das Verschulden des Beklagten vorausgesetzt. Da keine Pflicht zur proaktiven Kontrolle der Plattform seitens Provider besteht, liegt in der Regel kein Verschulden vor. Nach der Meldung eines illegalen Inhalts kann jedoch meist vernünftigerweise die Verhinderung oder die Beseitigung der Persönlichkeitsverletzung erwartet werden, womit das Verschulden zu bejahen wäre. Die Frage der erforderlichen Sorgfaltspflicht scheint hier jedoch noch weitgehend ungeklärt (m.w.H. BJ 2015: 99).

Unterliegt der Betreiber im Verfahren, so hat er die Verfahrenskosten zu tragen;[99] dies ungeachtet davon, ob ihm ein Verschulden vorgeworfen werden kann. In der Rechtslehre wird dies im Falle eines Betreibers, welchem der rechtsverletzende Inhalt nicht gemeldet wurde, als stossend angesehen (BJ 2015: 99).

Die Frage der Passivlegitimation und damit die allfällige Verantwortlichkeit von Internetprovidern wird im *Urheberrechtsgesetz* nicht geregelt. Die Rechtsprechung und die juristische Lehre gehen davon aus, dass für Provider eine Solidarhaftung nach Art. 50 Abs. 1 OR zu prüfen sei (m.w.H. BJ 2015: 40 f.). Schadenersatz wäre gemäss den Bestimmungen zur ausservertraglichen Haftung zu prüfen (Kuzniar 2017: 182).[100] Im Gegensatz dazu ist die *Stay-Down-Pflicht* im Urheberrechtsgesetz ausdrücklich verankert. Betreiber haben das erneute Aufschalten eines widerrechtlichen Inhalts auf ihrer Plattform zu verhindern. Diese Pflicht besteht, wenn ein Hinweis auf eine vorgängige Rechtsverletzung des Werks besteht und der Dienstanbieter durch seine Plattform die besondere Gefahr eines Missbrauchs ermöglicht.[101]

---

[98] BGE 126 III 161.

[99] Art. 106 ZPO.

[100] Art. 41 ff. OR.

[101] Art. 39d URG.

#### 4.2.3.3. Strafverfahren

Erfüllt ein Deepfake auch (mutmasslich) einen Straftatbestand, so kann allenfalls das strafrechtliche einem zivilrechtlichen Verfahren vorzuziehen sein (m.w.H. Rosenthal 2014: 417). Insbesondere hat das Strafverfahren den Vorteil, dass die Strafverfolgungsbehörden den Fall eigenständig ermitteln und Beweise vorbringen müssen und dabei auch mehr Möglichkeiten haben als eine Privatperson. Die Strafbehörde kann etwa bei Vorliegen eines strafrechtlich relevanten Verhaltens eine rückwirkende Teilnehmeridentifikation bewirken. Auch kann gegen einen anonymen Täter im Internet geklagt werden. Das Strafverfahren hat aber den Nachteil, dass die betroffene Person den Prozess nicht wirklich steuern kann (ebd.: 417). Für ein Verfahren im Zivilrecht spricht der grössere Schutzbereich der Ehre und auch das Vorliegen geringerer Entlastungsmöglichkeiten (Nobel/Weber 2021: 722). Im Strafrecht wird zudem vom Täter vorsätzliches Verhalten verlangt und es liegen kurze Verjährungsfristen vor. Die Art und Schwere der Sanktionen sind zudem unterschiedlich (Rosenthal 2014: 417; Zeller 2004: 199), und schliesslich kann eine strafrechtliche Untersuchung auch sehr viel länger dauern als ein zivilrechtliches Verfahren.

Zusätzlich zur Verfolgung von natürlichen Personen kann im Strafverfahren gegen eine *Plattform* vorgegangen werden. Werden gesetzlich verbotene Inhalte auf Plattformen verbreitet, die über einen Anknüpfungspunkt zur Schweiz verfügen, kann die Strafverfolgungsbehörde allenfalls auch eine Löschung anstreben. Ein Deepfake, welches als Beweismittel im Strafverfahren fungiert, könnte allenfalls beschlagnahmt werden.[102] Gewaltpropaganda, welche sich auf einem Schweizer Server befindet, kann mittels Verfügung des fedpols gelöscht werden.[103] Eine Löschung gestaltet sich schwierig, wenn der Hostingprovider den Sitz im Ausland hat. Falls sich das Material auf einem ausländischen Server befindet, kann das fedpol zudem den schweizerischen Providern die Sperrung der Website empfehlen (BAKOM 2011: 64 f.).[104]

Das öffentliche Interesse an anonymen Publikationen wird in der Schweiz anerkannt. So erlaubt das Strafgesetzbuch periodisch erscheinenden Medien, die

[102] Art. 263 StPO.

[103] Art. 13e des Bundesgesetzes über Massnahmen zur Wahrung der inneren Sicherheit (BWIS, SR 120).

[104] Art. 13e Abs. 5 BWIS.

Identität ihrer Autoren geheim zu halten.[105] Medien können die Herausgabe von IP-Adressen an Behörden verweigern. Blogbetreiber, die professionelle Medienschaffende sind, können sich ebenfalls auf diese Ausnahmebestimmung stützen. Bleibt der Autor unbekannt, hat die für die Veröffentlichung verantwortliche Person wegen Nichtverhinderns der strafbaren Veröffentlichung einzustehen (BAKOM 2011: 60).[106]

Betreiber von Plattformen, die nicht als Medienschaffende tätig sind, können zur Herausgabe von IP-Adressen verpflichtet werden. Ein Betreiber einer Internetplattform macht sich der Begünstigung strafbar, wenn er die IP-Adressen eines anonymen Autors von ehrverletzenden Kommentaren vernichtet (BAKOM 2011: 60; Robin/Christof 2022: 1249 ff.).[107]

Liegt eine klare Rechtsverletzung vor und reagiert der Provider trotz konkreten Hinweises auf illegalen Inhalt nicht, ist die Strafbarkeit aufgrund Gehilfenschaft zu überprüfen.[108] Inhaltsnahen Plattformen kann eine subsidiäre Strafbarkeit wegen Nichtverhinderns einer strafbaren Veröffentlichung zukommen.[109]

Welche dieser Bestimmungen für den jeweiligen Plattformbetreiber einschlägig ist, muss im Einzelfall betrachtet werden. Die strafrechtliche Verantwortlichkeit der Plattformen betreffend sämtliche illegale Inhalte erscheint in der heutigen Ausgestaltung des Internets nicht denkbar.[110]

#### 4.2.3.4. Zusammenfassung: Praktische Herausforderungen im Verfahren

Rechtlich gegen ein widerrechtliches Deepfake vorzugehen, ist nicht nur darum schwierig, weil derzeit noch einige Rechtsunsicherheiten bezüglich der Grenzen des Erlaubten bei Deepfakes bestehen, sondern weil das Verfahren mit einigen praktischen Herausforderungen verbunden ist. Schwierigkeiten bereiten namentlich die grenzüberschreitende Natur der Sachverhalte, die Identifizierung der Verantwortlichen, die internationale Zusammenarbeit im Rahmen

---

[105] Art. 28a StGB und Art. 172 StPO.

[106] Art. 322$^{bis}$ StGB.

[107] BGE 121 IV 109 E. 3.

[108] Art. 25 StGB.

[109] Art. 28 Abs. 2 und Art. 322$^{bis}$ StGB.

[110] Art. 305 StGB; BGE 121 IV 109 E. 3.

von Rechtshilfegesuchen sowie die Anerkennung und Durchsetzung schweizerischer Urteile im Ausland.

- *Grenzüberschreitende Sachverhalte:* Als problematisch erweist sich die Tatsache, dass Unternehmen von Social-Media-Plattformen ihren Geschäftssitz oftmals im Ausland haben und dass die sich dadurch ergebenden grenzüberschreitenden Sachverhalte zu Schwierigkeiten bei der Rechtsdurchsetzung führen. Handelt es sich um einen grenzüberschreitenden Sachverhalt, sind zunächst der Gerichtsstand und das anwendbare Recht zu bestimmen. Viele Plattformen verweisen in ihren Nutzungsbedingungen auf das ihren Geschäftssitz betreffende Recht und die staatlichen Gerichte vor Ort. Es ist jeweils im Einzelfall abzuklären, ob diese Klauseln standhalten; denn die zwingenden Bestimmungen des IPRG gehen vor. Das Bundesgericht entschied im Fall eines ehrverletzenden Eintrags auf Facebook, es sei gegen die Gesellschaft, welche die betreffende Seite betreibt, und nicht gegen den lokalen Ableger zu klagen. Nur diejenige Person, welche die Besitzerin der Daten sei oder zumindest die Kontrolle über die Daten habe, könne zu einer Herausgabe dieser Daten (Identität, Zugangsdaten, IP-Adresse des Kontoinhabers) verpflichtet werden.[111]
- *Identifizierung der Verantwortlichen:* Internetnutzer sind häufig anonym unterwegs. Gemäss Art. 22 Abs. 1 des BÜPF können Anbieterinnen von Fernmeldediensten bei Verdacht auf eine Straftat im Internet verpflichtet werden, dem Nachrichtendienst alle Angaben zu liefern, welche die Identifikation der Täterschaft ermöglichen. Das Strafverfahren muss dafür noch nicht eröffnet sein. Die Teilnehmeridentifikation und die Standortermittlung sind in der Strafprozessordnung geregelt.[112] Die Auskünfte können rückwirkend bis zu einer Frist von sechs Monaten verlangt werden. Eine entsprechende Anordnung benötigt die Genehmigung des Zwangsmassnahmengerichts. Die Anfrage zur Identifikation des Täters ist auch für die Verfolgung von Übertretungen zulässig. Auch die Polizei kann entsprechende Informationen einholen (Hansjakob 2018: Art. 22 N 1642). Bei einer Bedrohung der inneren oder äusseren Sicherheit kann der Nachrichtendienst auch präventiv die Datenherausgabe verlangen (ebd.: Art. 22 N 1644). Neben Anbieterinnen von Fernmeldediensten kommen auch Anbieterinnen von Diensten, die sich auf Fernmeldedienste stützen, gemäss BÜPF gewisse Auskunftspflichten

[111] BGE 143 IV 21.

[112] Art. 273 StPO.

zu. Aufgrund dieser vorgesehenen Massnahmen geht der Bundesrat davon aus, dass damit die Anonymität im Internet ausreichend aufgehoben werde könne und dass es keine zusätzliche zivilrechtliche Verantwortlichkeit der Intermediäre benötige (BJ 2015: 29).

- *Internationale Zusammenarbeit:* Liegt ein internationaler Sachverhalt vor, ist allenfalls zuerst abzuklären, welche Behörde für die Verfolgung der Straftat zuständig ist. Besteht ein allfälliger Kompetenzkonflikt zwischen der kantonalen Staatsanwaltschaft und der Bundestaatsanwaltschaft, kann Letztere ein Verfahren eröffnen.[113] Befinden sich die Betreiber der Plattform im Ausland, ist die Rechtsdurchsetzung schwierig. Die IP-Adressen befinden sich damit bei ausländischen Betreibern, welche nicht den Vorschriften des BÜPF unterworfen sind. Die Behörden sind auf die freiwillige Kooperation dieser Betreiber angewiesen. Manche Betreiber löschen zwar den Inhalt nicht, aber geben auf Anfrage von Strafverfolgungsbehörden IP-Adressen bekannt (BAKOM 2011: 60). Diese freiwillige Zusammenarbeit der Plattformbetreiber ermöglicht die Umgehung von aufwendigen Rechtshilfeverfahren (ebd.: 68; UVEK u.a. 2019: 33; BJ 2015). Kooperieren die Plattformen nicht, kann der Weg des internationalen Rechtshilfeverfahrens verfolgt werden. Aufgrund der hohen Zahl an grenzüberschreitenden Sachverhalten müssen die Ermittlungsbehörden häufig international tätig werden (BAKOM 2011: 66). Das Bundesgesetz über die internationale Rechtshilfe in Strafsachen (IRSG) gilt für Schweizer Behörden. Auf die zwischenstaatliche Zusammenarbeit besteht hingegen kein Anspruch, ausser es liegen spezifische Staatsverträge, wie das europäische Übereinkommen über die Rechtshilfe in Strafsachen, vor. Die Cybercrime Convention vom 1. Januar 2012 sieht zudem vor, dass Inhalte über Interpol oder Europol den zuständigen Behörden weitergeleitet werden. Diese entscheiden anhand des dort geltenden Rechts über die Löschung oder eine allfällige Strafverfolgung (ebd.: 66).
- *Anerkennung und Durchsetzung von Urteilen im Ausland:* Die Anerkennung eines Urteils eines zuständigen schweizerischen Gerichts im Ausland kann Probleme bereiten. Es ist nicht garantiert, dass ein entsprechendes Gerichtsurteil im Ausland tatsächlich umgesetzt wird (BAKOM 2011: 65). Grundsätzlich ist die Anerkennung sowie die Vollstreckung eines Urteils in der Schweiz möglich. Dafür erforderlich ist das Durchlaufen eines Anerkennungs- und Vollstreckungsverfahrens vor den zuständigen Gerichten im Ausland. Hier kann die Gegenpartei in einem erneuten Gerichtsverfahren die Umsetzung

[113] Art. 27 Abs. 2 StPO.

verzögern oder gar verhindern, etwa durch das Vorbringen von abweichenden Bestimmungen nach ausländischem Recht. Liegt kein internationales Übereinkommen über die gegenseitige Anerkennung vor, wird das nationale Gericht die Rechtswidrigkeit nach dort herrschendem Recht prüfen. Die Mitgliedstaaten des LugÜ erkennen schweizerische Urteile an. Zudem benötigt es kein separates Verfahren für eine Vollsteckbarkeitserklärung. Vielmehr kann sich die betroffene Person direkt an die Vollstreckungsbehörde des betreffenden Staates wenden. Ob eine Klage in der Schweiz sinnvoll erscheint, ist also davon abhängig, wo das Urteil zu vollstrecken sein wird und ob dieser Staat das Urteil anerkennen wird (BJ 2015: 97). Je nach Fall ist auch eine Geltendmachung der Ansprüche direkt vor Ort zu erwägen. Die von einem illegalen Deepfake betroffene Person kann sich bei Vorliegen einer strafbaren Handlung direkt an die zuständige ausländische Strafverfolgungsbehörde wenden. Auch allfällige zivilrechtliche Ansprüche können direkt vor Ort eingeklagt werden. Diesem Vorgehen stehen aber bedeutende praktische und monetäre Hürden im Wege, so etwa die Tatsache, dass derartige Verfahren nicht ohne anwaltliche Vertretung im anderen Staat geführt werden können (Rosenthal 2014: 418).

### 4.2.4. Zwischenfazit zum Schutz vor Deepfakes

Deepfakes sind in der Schweiz nicht ausdrücklich reguliert. Als Folge davon muss auf allgemeine zivil- und strafrechtliche Bestimmungen zurückgegriffen werden, um gegen schädliche Deepfakes vorzugehen. Viele der «unerwünschten» Anwendungsfälle von Deepfakes dürften durch das geltende Recht erfasst werden.[114]

Jedoch gibt es Graubereiche, namentlich bei der strafrechtlichen Beurteilung von «peinlichen» oder «freizügigen» Deepfakes. Das existierende Strafrecht vermag zudem den besonderen Unrechtsgehalt von pornografischen Deepfakes nicht immer zu erfassen, namentlich wenn diese gleichzeitig eine Verletzung der Ehre als auch der sexuellen Selbstbestimmung beinhalten.

Die grössten Herausforderungen zeigen sich im Bereich der Durchsetzung zivil- und strafrechtlicher Ansprüche. Im Internet sind die Täter häufig unbekannt oder mit Sitz im Ausland. Auch ein Vorgehen gegen Plattformen ist mit prozessualen

[114] So auch die Stellungnahme des Bundesrats auf die Motion 23.3563 Mahaim «Deepfakes regulieren» vom 16.08.2023.

Unsicherheiten behaftet. Schliesslich können insbesondere im Zivilverfahren die Verfahrenskosten auf Betroffene von Deepfakes abschreckend wirken.

## 4.3. Deepfakes vor Gericht

Gerichte und Gerichtsverfahren stellen einen potenziellen Einsatzkontext von Deepfakes dar. Deepfakes können beispielsweise verwendet werden, um vor Gericht verwendete Beweismittel zu manipulieren. Dies wirft zivil- und strafprozessuale Fragen auf (Kapitel 4.3.1). Deepfakes könnten aber auch zu «nicht schädlichen» forensischen Zwecken verwendet werden, etwa zur Visualisierung von Tathergängen oder im Rahmen von verdeckten Ermittlungen. Deren Zulässigkeit wirft ebenfalls rechtliche Fragen auf (Kapitel 4.3.2).

### 4.3.1. Deepfakes als (manipulierte) Beweismittel

Mittels Deepfakes können täuschend echte Bild-, Video- oder Audiodateien kreiert werden, die auch als Beweismittel in Gerichtsverfahren zum Einsatz kommen könnten. So könnte mittels eines Deepfakes ein falsches Alibi kreiert werden. Dabei ist zu beachten, dass alle mechanisch hergestellten Dokumente dem Risiko einer Verfälschung unterliegen (Hasenböhler 2019: 195). Das Einbringen von gefälschten Urkunden vor Gericht ist insofern keine Neuigkeit per se. Programme wie Photoshop ermöglichen bereits seit vielen Jahren die Manipulation von Texten oder Fotos (Pfefferkorn 2020: 271). Deepfake-Technologien sind eine neue technische Anwendung, die das Fälschen von Video und Ton erleichtern, aber nicht grundsätzliche neue Fragen aufwerfen.

Das *Zivilverfahren* wird in der Zivilprozessordnung (ZPO)[115] geregelt. Die zulässigen Beweismittel werden in Art. 168 ZPO aufgeführt. Als Urkunden gelten neben klassischen Schriftstücken wie Aufzeichnungen auf Papier auch Fotos, Filme und Tonaufzeichnungen sowie sonstige elektronische Dateien.[116]

Gemäss Art. 178 ZPO wird die Echtheit von eingeführten Urkunden vermutet. Wird sie von der Gegenpartei mit einer ausreichenden Begründung bestritten, muss diejenige Partei, welche sich auf die Urkunde beruft, deren Echtheit be-

[115] Schweizerische Zivilprozessordnung (ZPO) vom 19. Dezember 2008, SR 272.

[116] Art. 177 ZPO.

weisen (Hasenböhler 2019: 208 f.). Gemäss Bundesgericht geht es bei dieser Bestimmung lediglich um die Echtheit im engeren Sinn, also um die Frage, ob die Urkunde wirklich von derjenigen Person stammt, die als Urheber erkennbar ist, und nicht um die inhaltliche Richtigkeit des Dokuments (ebd.: 207).[117]

Im Rahmen der Beweiswürdigung beurteilt das Gericht, welche Beweiskraft einem Beweis zukommt (Hasenböhler 2019: 195). Die ZPO erkennt alle Beweismittel als gleichwertig an; in der Praxis wird Urkunden ein vergleichsweise hoher Beweiswert zugesprochen. Das Gericht hat aber den Einzelfall nach dem Prinzip der freien Beweiswürdigung nach Art. 157 ZPO vorzunehmen und über die Authentizität der Urkunde zu entscheiden (m.w.H. Hasenböhler 2019: 213f.). Gemäss Art. 183 Abs. 1 ZPO kann das Gericht auch von Amtes wegen entsprechende Gutachten einholen,[118] was bei dem Einwand, es handle sich um einen Deepfake, in Zukunft häufiger der Fall sein dürfte (Kuhlmann 2020).

Im *Strafverfahrensrecht* hingegen gilt die Unschuldsvermutung «in dubio pro reo» als Beweiswürdigungsregel. Es darf nicht vom für den Angeklagten ungünstigeren Sachverhalt ausgegangen werden, wenn objektiv betrachtet Zweifel bestehen, ob sich der Sachverhalt entsprechend zugetragen hat (Groner 2011: 164).[119] Diese Beweiswürdigungsregel könnte auch im Falle von Deepfakes bedeutend sein.

Der Einwand, eine Urkunde sei ein Deepfake, könnte auch als Verfahrenstaktik angewendet werden. Es könnte aber auch vorkommen, dass ein Beweis eingeführt wird, ohne dass sich die Partei selbst bewusst ist, dass es sich um ein Fake handelt (Pfefferkorn 2020: 255; van der Sloot/Wagensveld 2021; van Huijstee u.a. 2021: 55).

Aus der Generalklausel von Art. 12 lit. a des Anwaltsgesetzes,[120] wonach Anwältinnen und Anwälte ihren Beruf sorgfältig und gewissenhaft auszuführen haben, leitet die Rechtslehre ab, dass es unzulässig sei, das Gericht durch die Einreichung unrichtiger Beweismittel über die Beurteilung eines wesentlichen Sachverhalts in die Irre zu führen (Fellmann 2011: N 37). Wird absichtlich ein Deepfake in den Gerichtsprozess eingespielt oder ein Video wider besseren Wissens

---

[117] BGE 143 III 453 E. 3.

[118] Bei unbeschränktem Untersuchungsgrundsatz auch allenfalls dazu verpflichtet (Hasenböhler 2019: 300 f.).

[119] Art. 10 Abs. 3 StPO; Art. 32 Ab. 1 BV und Art. 6 Ziff. 2 EMRK.

[120] Bundesgesetz über die Freizügigkeit der Anwältinnen und Anwälte (BFA, SR 935.61).

als Deepfake bezeichnet, wäre eine Verletzung des Standesrechts durch den Anwalt zu prüfen (Pfefferkorn 2020: 274).

Weiter ist darauf hinzuweisen, dass die Fälschung einer Urkunde eine Straftat darstellt.[121] Wer jemanden wider besseren Wissens wegen einer Straftat bei einer Strafbehörde anzeigt, kann sich, aufgrund falscher Anschuldigung[122] resp. Irreführung der Rechtspflege,[123] schuldig machen.

### 4.3.2. Einsatz von Deepfakes zur Aufklärung von Straftaten

Denkbar ist, dass Deepfake-Technologien von den Strafverfolgungsbehörden genutzt werden, um Straftaten aufzuklären. So kann beispielsweise mittels Deepfake-Technologie aus Handyvideos, Überwachungskameras sowie Körperscans ein virtueller Tatort kreiert oder ein *Tathergang rekonstruiert werden*. Neben epistemologischen Fragen (etwa der Gefahr des Anscheins von objektiver Wahrheit, während die Herstellung einseitig durch die Strafverfolger erfolgt) existieren auch rechtliche Unsicherheiten, namentlich die Frage, wie die Gewährung der Teilnahmerechte in allen Verfahrensstadien und somit auch bei der «virtuellen Tatortbegehung») sichergestellt werden kann, die Frage der Überprüfbarkeit der derart erhobenen Beweise oder die Frage, wie und in welcher Form digital erhobene Beweise künftig zu den Akten gelegt werden oder inwiefern Beteiligte zur Teilnahme an einer virtuellen Rekonstruktion verpflichtet sind (Gjon 2019: 25). Rechtsprechung oder Lehrmeinungen zu diesen Fragen fehlen derzeit noch weitgehend.

Ebenfalls diskutiert wird, ob verdeckte Ermittler *computergenerierte Kinderpornografie verwenden dürfen, um Zutritt zu Onlineforen zu erhalten*, die ein Hochladen von eigenen Bildern oder Videos voraussetzen (Menkens 2019). Da auch das Verbreiten von «fiktiver» Kinderpornografie eine Straftat ist (s.o. 4.2.2.6), liegt die Problematik darin, dass nach schweizerischem Strafprozessrecht verdeckte Ermittler im Rahmen ihres Einsatzes keine Straftaten begehen dürfen.[124]

---

[121] Art. 251 StGB.

[122] Art. 303 StGB.

[123] Art. 304 StGB.

[124] Gemäss Art. 288 Abs. 3 StPO wird bei Begehen einer Straftat durch einen verdeckten Ermittler ein Strafverfahren durchgeführt, was impliziert, dass grundsätzlich ein gesetzeskonformes Verhalten der verdeckten Ermittler erwartet wird.

In der Rechtslehre wird die Ansicht geäussert, dass solches Verhalten juristisch gerechtfertigt sein könnte, wenn das Anbieten von verbotenem Bildmaterial durch einen verdeckten Ermittler als eine Demonstration der wirtschaftlichen Leistungsfähigkeit interpretiert wird, die nach Art. 293 Abs. 3 StPO explizit zulässig ist. Die Leistungsfähigkeit im Pornografiegeschäft werde im Gegensatz zu anderen Deliktsbereichen regelmässig nicht durch Geld demonstriert, sondern durch entsprechendes Bildmaterial (Muggli 2014: 305). Zudem stellt die Verwendung von fiktivem Material – sollte solches in überzeugender Qualität generiert werden können – zumindest moralisch bzw. aus Sicht des Opferschutzes eine mildere Massnahme dar als die Verwendung von echtem Material, das den Ermittlern aus polizeilichen Beschlagnahmungen zur Verfügung stehen würde. Gegen dieses Argument wird jedoch ins Feld geführt, dass «täuschend» echte Deepfakes ohne «echtes» Bildmaterial derzeit technisch nicht erstellt werden können, sodass in jedem Fall ein Kindesmissbrauch stattgefunden haben muss (etwa für die Trainingsdaten). Richterliche Praxis zu dieser Frage existiert derzeit noch nicht.

### 4.3.3. Zwischenfazit

Deepfakes werden voraussichtlich auch in zunehmendem Masse im Rechtsalltag eine Rolle spielen. Hier untersucht wurden die Einsatzszenarien von Deepfakes als manipulierte Beweismittel sowie von Deepfakes als Hilfsmittel zur Aufklärung von Straftaten. Während das Prozessrecht grundsätzlich Antworten für den Umgang mit manipulierten oder gefälschten Beweismitteln bereithält, stellen sich im Hinblick auf den Einsatz von Deepfakes zur Aufklärung von Straftaten noch ungeklärte prozessrechtliche Fragen, etwa die praktische Wahrung der Beschuldigtenrechte oder die Zulässigkeit der Verwendung von mittels Deepfake-Technologie erstellter «Kinderpornografie» durch die Polizei zwecks Eindringens in geschlossene Foren zur Ermittlung. Viele diese Fragen stellen sich nicht nur bei Deepfakes, sondern allgemein bei digitalen Beweismitteln im Strafverfahren.

## 4.4. Öffentlich-rechtliche Vorgaben

Deepfakes können zur Desinformation und Manipulation der Bevölkerung genutzt werden. Deepfakes, welche Politiker blossstellen, Geschehnisse vortäuschen oder Militäranschläge simulieren, könnten im schlimmsten Fall Massendemonstrationen bewirken oder sogar eine Regierungs- und Staatskrise

auslösen, zu einer Wahlbeeinflussung oder auch zu einem Einbruch der Aktienmärkte führen (van Huijstee u.a. 2021; Lantwin 2019: 574 ff.) (vgl. auch Kapitel 6 und 7).

Dem Staat kommt die Aufgabe zu, die (Kommunikations-)Grundrechte zu verwirklichen. Entsprechend sind die notwendigen Rahmenbedingungen gesetzlich festzulegen (Hertig 2015b: 216). Unter Umständen kann für den Erhalt einer funktionierenden Demokratie und zum Schutz pluralistischer Informationen und Meinungsbildung eine staatliche Intervention geboten sein.[125] Nachfolgend werden die bestehenden Rechtsvorgaben im Bereich der Medienregulierung sowie zum Schutz von Wahlen und Abstimmungen dargestellt.

### 4.4.1. Medienregulierung

Der Schutz der unverfälschten Meinungs- und Willensbildung der Öffentlichkeit ist in der Bundesverfassung reguliert. Die Vorschrift betrifft aber nur die audiovisuellen Medien. Radio und Fernsehen wird nämlich im Vergleich zu den Printmedien eine unmittelbarere und stärkere Wirkung zugesprochen und damit ein grösseres Beeinflussungspotenzial und Anfälligkeit zur Manipulation (m.w.H. Hertig 2015b: 252).[126] Gemäss der Verfassungsbestimmung haben Radio und Fernsehen Ereignisse sachgerecht darzustellen und die Meinungsvielfalt zu repräsentieren (ebd.: 252; Zeller 2004: 237).[127] Das in Art. 4 Abs. 2 RTVG festgehaltene Sachgerechtigkeitsgebot bezweckt keine Wahrheitspflicht, soll aber als journalistische Sorgfaltspflicht verstanden werden, indem etwa Berichterstattungen transparent zu halten sind, um eine Nachvollziehbarkeit zu gewähren (Cueni: 3 ff.). Allfällige Beschwerden gegen Radio- und Fernsehprogramme können bei der unabhängigen Beschwerdeinstanz (UBI) eingereicht werden.[128]

Internetanbieter, welche selbst Inhalte publizieren, müssen sich an gewisse gesetzliche Regulierungen halten. Keinen Vorschriften unterliegen Internetdienste (Social-Media-Kanäle/Privat), die keinen eigentlichen Programmcharakter haben. Da Publikationen im Internet häufig keiner Strukturierung und Moderation unterstehen und auch das Gebot der sachgerechten Darstellung nicht einzu-

---

[125] BGE 120 Ib 142 E. 4b/aa.

[126] EGMR, Murphy c. Irland, Nr. 44179/98 (2003). BGE 98 Ia 73, 91 E. 3c; Brunner/Burkert, St. Galler Kommentar zu Art. 17 N 8.

[127] Art. 93 Abs. 2; BGE 134 I 2 E. 3 S. 5 ff.

[128] Art. 93 Abs. 5 BV.

halten haben und zugleich von tiefen Kosten und dem grossen Verbreitungspotenzial profitieren, kann dies zum raschen Verbreiten von Des- und Misinformationen führen.

Anwendungen von Deepfakes könnten auch aufgrund ihrer audiovisuellen Darstellung eine verstärkte Glaubwürdigkeit mit Manipulationspotenzial implizieren (Lantwin 2019: 574 f.; Meinicke 2020: 981 ff.). Weiter generieren Des- und Misinformation häufig mehr Aufmerksamkeit als die anschliessende Richtigstellung, weshalb ein Teil der rezipierenden Personen weiterhin falsch informiert bleiben könnte. Personen, die einmal durch ein Deepfake getäuscht wurden, könnten in Zukunft auch wahrheitsgetreue Informationen als Deepfake aufnehmen (van der Sloot/Wagensveld 2021: 4). Beim Einzelnen kann dies zu einem Untergraben des Vertrauens in den öffentlichen Diskurs führen, was sich auch auf gesellschaftlicher Ebene auswirken könnte (van Huijstee u.a. 2021: 4).

Während bei der Verbreitung von Deepfakes über Radio und TV transparent darauf hingewiesen werden müsste (BAKOM 2011: 14), besteht für Social-Media-Plattformen und Intermediäre keine Pflicht zur Moderation dieser. Zudem kann der Staat – aufgrund der zu schützenden Meinungsfreiheit – nur sehr zurückhaltend regulieren und nur gegen offensichtlich rechtswidrige Inhalte vorgehen (Tschannen 2021: 400 N 1051). Die Selbstregulierung von Plattformen wird weiter unten (vgl. Kapitel 4.5.2.4) besprochen.

Wenig Handhabe besteht innerstaatlich gegen von ausländischen Medien – etwa zu Propagandazwecken – verbreitete Deepfakes. Das RTVG erstreckt sich nur auf Schweizer Veranstalter. Das Europäische Übereinkommen über das grenzüberschreitende Fernsehen (EÜGF; SR 0.784.405) folgt dem Sendestaatsprinzip, sieht aber, wie das RTVG, keine expliziten Bestimmungen gegen Deepfakes oder Des- und Misinformation vor. Nur wenn die (auch für die Schweiz verbindlichen) völkerrechtlichen Vorschriften über Programmgestaltung, Werbung oder Sponsoring dauernd und schwerwiegend verletzt sind, könnte das BAKOM ein ausländisches Programm einschränken. Bei Programmen, die aus Nicht-EÜGF-Staaten gesendet werden (dazu gehören auch etwa Russland oder die USA), ist die Schweiz nicht an die Vorgaben des EÜGF gebunden.

Hinzuweisen ist ferner auf das Bundesgesetz über den Jugendschutz in den Bereichen Film und Videospiele (JSFVG, BBl 2022 2406). Das Gesetz regelt nicht primär Deepfakes, sondern behandelt v.a. Darstellungen von Gewalt, Sexualität und bedrohlichen Szenen, in deren Kontext aber durchaus auch Deepfakes denkbar wären. Das Gesetz verpflichtet die entsprechenden Verbreiter (z.B. Detailhandel, Streamingdienste, Kinounternehmen etc.) zu bestimmten

Jugendschutzmassnahmen wie z.B. Alterskennzeichnung und -kontrollen und ggf. Zugangsbeschränkungen. Die Umsetzung der Massnahmen geschieht im Rahmen einer Koregulierung.

### 4.4.2. Schutz von Wahlen und Abstimmungen vor Verfälschung

Für Wahlen und Abstimmungen gelten besondere Garantien. Die Bundesverfassung schützt die freie Willensbildung und die politischen Rechte.[129] Stimmbürger haben einen Anspruch auf die rechtzeitige Publikation der Abstimmungsvorlagen und auf den rechtzeitigen Versand des Stimmmaterials. Weiter kommt dem Staat eine verfassungsrechtliche Informationspflicht zu.[130] Behörden dürfen grundsätzlich in Wahlkämpfen nicht intervenieren. Ausnahmsweise sind punktuelle Eingriffe in politische Debatten zulässig, um irreführende Informationen richtigzustellen. Es soll eine ausgewogene Information bereitgestellt werden, um eine unverfälschte Willensbildung zu ermöglichen.

Private können jedoch im Wahlkampf jederzeit intervenieren. Lediglich irreführende Angaben unmittelbar vor dem Urnengang sind unzulässig, weil damit die unverfälschte Meinungsbildung verwehrt würde (Tschannen 2021: 708 N 1909).[131] Eine Wahl oder ein Abstimmungsergebnis nach einer solch kurzfristigen privaten Intervention mit grosser Tragweite darf aufgrund der verfälschten Willensbildung nicht anerkannt werden. Im Unterlassensfall kann Beschwerde gegen das Gemeinwesen erhoben werden (ebd.: 708 N 1909).

Das hier aufgezeigte Vorgehen während Wahlen und Abstimmungen wäre auch bei der Verbreitung von entsprechenden Deepfakes einschlägig. Die Behörde hätte bei einer Verbreitung eines Deepfakes mit irreführenden Angaben zu intervenieren und die Sachlage richtigzustellen (Tschannen 2021: 702 N 1894). Ist eine Richtigstellung zeitlich nicht mehr möglich oder wird diese durch das Gemeinwesen unterlassen, wäre das Ergebnis im Anschluss nicht anzuerkennen.

Denkbar wäre auch, dass ein anderer Staat oder terroristische Organisation gezielte Desinformation betreibt, um Unruhen oder ein gewünschtes Wahlergebnis zu erzielen (siehe hierzu van der Sloot/Wagensveld 2021: 3). Dies würde gegen das völkerrechtliche Interventionsverbot verstossen. Die Schwierigkeit hierbei

---

[129] Art. 34 Abs. 2 BV.

[130] BGE 132 I 104 E. 3.2, S. 108.

[131] BGE 119 Ia 271 E. 3c, S. 274; BGE 135 I 292 E. 4.1, S. 295.

wäre, dass die nationale Gesetzgebung gegenüber einem anderen Staat nicht durchgesetzt werden kann und bereits die Identifikation der Täter nur schwer möglich wäre (van Huijstee u.a. 2021: 53).

So hat der Europarat im Übereinkommen über Computerkriminalität (SEV Nr. 185) vom 23. November 2001 anerkannt, dass neben Cyberbedrohungen mit physischen und ökonomischen Konsequenzen auch Desinformationen, welche dem demokratischen Prozess schaden könnten, als Cyberattacken anzusehen sind (Doublet 2019: 21).

Die Wahrnehmung der politischen Autonomie ist vom Einzelnen, der Gesellschaft und den wirtschaftlichen Rahmenbedingungen abhängig. Gemäss Tschannen habe der Staat auf diese Faktoren nur wenig Einfluss. Der Staat könne aber durch Bildung, Sicherung eines vielfältigen Mediensystems, Gewährleistung von Kommunikationsgrundrechten und Einrichtung politischer Partizipationsmöglichkeiten den Rahmen für die Wahrnehmung der politischen Autonomie zur Verfügung stellen (Tschannen 2021: 401 N 1054). Bund und Kantone haben sich dafür einzusetzen, dass Kinder und Jugendliche in ihrer Entwicklung zu selbstständigen und sozial verantwortlichen Personen gefördert und ihre soziale, kulturelle und politische Integration unterstützt werden. Bildung wird als Schlüssel für das erfolgreiche Zusammenleben in der Gesellschaft angesehen. Das Fach Medien und Informatik des Lehrplans 21 soll Digital- und Medienkompetenzen der Schülerinnen und Schüler stärken. Diese sollen unter anderem die Fähigkeit erwerben, Informationen aus verschiedenen Quellen gezielt zu beschaffen, auszuwählen und hinsichtlich Qualität und Nutzen zu beurteilen.[132] Auch der Europarat fordert die Bereitstellung der Mittel zur Förderung der Medien- und Digitalkompetenz (Council of Europe 2016).

### 4.4.3. Zwischenfazit

Der Gesetzgeber hat im öffentlich-rechtlichen Bereich, namentlich in der Medienregulierung sowie beim Schutz politischer Prozesse vor Beeinflussung, Regelungen gegen das Verbreiten falscher Informationen erlassen. Jedoch stammen viele dieser Regelungen noch aus dem «analogen» Zeitalter und erfassen die technologiebedingten Besonderheiten (grenzüberschreitende Verbreitung, Schnelligkeit der Verbreitung, Unklarheit der Urheberschaft) noch nicht in genügendem

[132] Siehe auch Lehrplan 21 (2016), BAKOM (2018) sowie Jugend und Medien (2023).

Masse. Weitere Regulierungsmöglichkeiten werden deshalb im nachfolgenden Kapitel dargestellt und fliessen in die Empfehlungen dieses Berichts ein.

## 4.5. Regulierungsmöglichkeiten von Deepfakes

Im Folgenden geht es darum zu untersuchen, ob und inwiefern die bestehende Gesetzgebung zum Umgang mit Deepfakes ausreichend ist und inwieweit weiterer Regulierungsbedarf besteht. Dazu wird zunächst auf die bereits identifizierten sowie einige weitere Herausforderungen von Deepfakes, die möglicherweise einen Regulierungsbedarf begründen, eingegangen. In einem zweiten Schritt werden bestehende Lösungsansätze, wie sie entweder in anderen Staaten bereits umgesetzt wurden oder in der Literatur gefordert werden, aus einer rechtlichen Sicht diskutiert. Konkrete Regulierungsvorschläge für die Schweiz finden sich im Kapitel zu den Empfehlungen am Ende dieser Studie (vgl. Kapitel 8).

### 4.5.1. Allgemeine Herausforderungen von Deepfakes

Die juristische Untersuchung hat gezeigt, dass die schädlichen Anwendungen von Deepfakes bis auf einige Grenzbereiche weitestgehend vom materiellen Recht abgedeckt sind. Faktisch stellen sich die grössten rechtlichen Herausforderungen bei der Durchsetzung von Ansprüchen in Zivil- oder Strafverfahren.

Allerdings ergeben sich auch allgemeine, strukturelle Herausforderungen, die durch eine isolierte Untersuchung von juristischen Fragestellungen nicht zutage treten, die aber für die Begründung eines Regulierungsbedarfs ebenfalls relevant sind. Sie werden im Folgenden kurz beschrieben.

#### 4.5.1.1. Vertrauen in die Rechtsinstitutionen

Eine in Bezug auf Deepfakes verbreitete Sorge ist, dass die weite Verbreitung von Deepfake-Videos zu grundlegenden Vertrauensproblemen in der Gesellschaft führen könnte. Wahre Informationen, die durch Behörden mitgeteilt werden, könnten als verfälscht aufgenommen und schliesslich zu einem Vertrauensverlust der Bevölkerung in den demokratischen Rechtsstaat und in die Justiz führen (van Huijstee u.a. 2021: 55). Postuliert wird auch die Gefahr, dass die Bevölkerung nicht mehr an die Möglichkeit der Wahrheitsfindung in einem Gerichtsprozess glaubt (Pfefferkorn 2020: 255). Verurteilte Personen könnten

weiterhin öffentlich an ihrer Unschuld festhalten und dem zuständigen Gericht einen Fehlentscheid vorwerfen. Sie könnten etwa behaupten, dass es sich bei den verwerteten Beweisen um absichtlich manipulierte Urkunden handle. Umgekehrt könnte beispielsweise ein gefälschtes Video – welches jemanden bei einer Straftat zeigt – aufgrund seiner Verbreitung auch nach Freispruch an der Person haften bleiben (van der Sloot/Wagensveld 2021: 3). Der damit anhaltende Vorwurf der Straftat könnte zu Auswirkungen auf das Privat- und Berufsleben einer Person im Sinne eines Reputationsschadens führen (van der Sloot/Wagensveld 2022: 6).

#### 4.5.1.2. Geschlechtsspezifische Gewalt

Untersuchungen zeigen, dass das Phänomen illegaler Deepfakes hauptsächlich die pornografische Darstellung von Frauen betrifft (van der Wilk 2021: 28). Gemäss der General Recommendation von GREVIO, dem Überwachungsorgan zur von der Schweiz ebenfalls ratifizierten Istanbul-Konvention (Council of Europe 2011; GREVIO 2021), sind manipulierte pornografische Bilder einer Person (GREVIO 2021: 18) eine Form der sexuellen Belästigung. Die Vertragsparteien sind nach Art. 40 der Istanbul-Konvention verpflichtet, die erforderlichen gesetzgeberischen oder sonstigen Massnahmen zu treffen, um dieses Verhalten zu sanktionieren. Das Phänomen von Revenge Porn, bei welchem private Aufnahmen von intimen Akten – meistens von Ex-Partnern – veröffentlicht werden, könnte im Zusammenhang mit Deepfake-Technologien verstärkt auftreten (van der Sloot/Wagensveld 2021: 3). Ungeachtet der Echtheit solcher Videos kann dies die psychische Gesundheit betroffener Person schwer beeinträchtigen (m.w.H. ebd.: 3 und Šepec 2020: 429). In der Schweiz wäre eine vertiefte Diskussion zu begrüssen zur Klärung der Frage, ob diese Verletzungen anstelle von Ehrverletzungsdelikten zu den Straftaten gegen die sexuelle Integrität zugeordnet werden sollten (siehe Šepec 2020), denn bildbasierte sexuelle Gewalt verletzt nicht nur den Bereich der Ehre, sondern auch der sexuellen Selbstbestimmung. Der Bundesrat hat im Jahr 2022 einen Bericht zum Thema Cybermobbing veröffentlicht, in welchem er zum Schluss kam, dass derzeit kein Handlungsbedarf im materiellen wie im Verfahrensrecht besteht (Bundesrat 2021). Andere Staaten haben derweil hingegen spezifische Straftatbestände geschaffen, die den bildbasierten sexuellen Missbrauch sanktionieren (van der Wilk 2021: 57). Auch stellt sich die Frage, ob das Erstellen von Deepfakes mit pornografischem Inhalt – etwa von dem Ex-Freund oder der -Freundin – auch für den Privatgebrauch kriminalisiert werden sollte (Šepec 2020: 428; van der Sloot/Wagensveld 2022: 10).

#### 4.5.1.3. Social cooling

Die vermehrte Verbreitung von Deepfakes führt möglicherweise zu einem *social cooling*, eine Einschränkung der Veröffentlichung von Missständen aufgrund der Angst von Betroffenen, anschliessend Opfer von Deepfakes zu werden (van Huijstee u.a. 2021: 30).

#### 4.5.1.4. Zugang zum Recht

Bereits die Entscheidung, einen Prozess zu führen, kann Privaten – insbesondere mit im Zivilverfahren entstehenden Kosten und Zeitaufwand – Hürden stellen. Ein Rechtsstreit kann zudem unerwünschte Aufmerksamkeit auf das Streitobjekt lenken (van der Sloot/Wagensveld 2022: 12). Ein weiteres Problem für die Rechtsdurchsetzung stellt die allfällige erneute Weiterverbreitung veröffentlichter Inhalte durch andere Internetnutzende dar, was es schwierig bis unmöglich machen kann, die rechtlichen Ansprüche durchzusetzen (ebd.: 10; BAKOM 2011: 69 f.; van Huijstee u.a. 2021: 50).

Das Problem der internationalen Rechtsdurchsetzung ist ein allgemeines Problem und betrifft nicht nur spezifisch illegale Deepfakes. Verfahren, die Straftaten im Internet betreffen, gelten als sehr ressourcenaufwendige Verfahren. Häufig sind mehrere Personen involviert, welche identifiziert werden müssten, oder es stellt sich gar das Problem der Unmöglichkeit einer Identifikation (van der Wilk 2021: 12; van Huijstee u.a. 2021: 53). Unklare Zuständigkeiten sowie die Überlastung von Strafverfolgungsbehörden kommen dazu (van der Sloot/Wagensveld 2022: 12; van der Wilk 2021: 12). Rechtshilfeabkommen als auch vertiefte Kooperationen im Bereich des Datenaustauschs werden als nützliche Instrumente anerkannt, setzen aber grosses Vertrauen zwischen den Vertragsstaaten voraus (BJ 2015: 4; BAKOM 2011: 72 f.).

Mit dem Inkrafttreten des totalrevidierten DSG haben private Datenbearbeiter mit Sitz im Ausland die Pflicht, eine Vertretung in der Schweiz zu benennen.[133] Dies soll eine erleichterte Kontaktaufnahme mit den Bertreibern von Internetplattformen bezwecken. Dadurch wird jedoch keine zwangsweise Rechtsdurchsetzung ermöglicht (siehe auch van Huijstee u.a. 2021: 61). Der Digital Service Act sieht stattdessen eine Verantwortlichkeit dieser Rechtsvertreter vor (siehe dazu nachfolgend Kapitel 4.5.2.2).

[133] Art. 14 DSG.

### 4.5.2. Bestehende Regulierungsansätze

Nachfolgend werden Regulierungsansätze, wie sie bereits praktiziert werden, aus einer rechtlichen Sicht dargestellt.

Generell können drei Ansätze zum Umgang mit Deepfakes unterschieden werden: Der erste besteht darin, Gesetzgebung spezifisch zur Bekämpfung bzw. Regulierung von Deepfakes zu adoptieren. Verschiedene US-Gliedstaaten sowie der amerikanische Bundesstaat haben diesen Ansatz gewählt (s.u. Kapitel 4.5.2.1) (Jacquemin 2023: 324 f.). Der zweite Ansatz reguliert mit einem technikneutralen Ansatz zwar nicht Deepfake-Technologien als solche, jedoch in einem breiteren Sinn die dadurch verursachten Probleme, so etwa die KI-Regulierung in der EU (s.u., Kapitel 4.5.2.2) oder das Netzwerkdurchsetzungsgesetz in Deutschland (s.u., Kapitel 4.5.2.3). Der dritte Ansatz schliesslich besteht darin, keine spezifische Regelung anzunehmen in der Annahme, dass ein Abstützen auf allgemeine Bestimmungen ausreicht, um dem Phänomen zu begegnen. Dies ist der bisher in der Schweiz gewählte Ansatz, wobei dieser nicht auf eine aktive gesetzgeberische Entscheidung, sondern hauptsächlich auf ein Abwarten zurückzuführen ist.

#### 4.5.2.1. USA

In den USA haben verschiedene Bundesstaaten spezifische Regulierungen für Deepfakes erlassen. Dabei lassen sich jene im Hinblick auf intime Darstellungen von jenen mit Ziel der Wahlmanipulation unterscheiden. So verbietet etwa Kalifornien innerhalb 60 Tagen vor einer Wahl die Manipulation politischer Kandidaten, welche den Betroffenen schaden sollen. Diese Regelung stand in der Kritik wegen der dadurch entstehenden Einschränkung der Meinungsfreiheit; auch weil keine Ausnahmen für Satire oder Parodie vorgesehen sind. Zudem wurden Regeln vorgesehen, um das Beschwerdeverfahren für Einzelpersonen bei pornografischen Deepfakes zu vereinfachen (m.w.H. van Huijstee u.a. 2021: 45; Feeney 2021 und Kirchengast 2020: 308 ff.). Ein ähnliches Verbot betreffend Erstellung und Weiterverbreitung pornografischer Deepfakes besteht in New York, welches zudem als postmortales Recht Künstlerinnen und Künstler während 40 Jahren nach Ableben vor Deepfakes zur kommerziellen Verwendung schützt (The New York State Senate 2020).

Der vom amerikanischen Kongress verabschiedete Deepfake Accountability Act[134] sieht u.a. vor, dass Deepfakes mit einem Wasserzeichen markiert werden müssen (Hewage 2020).

#### 4.5.2.2. Europäische Union[135]

**Verordnung der EU über digitale Dienste (DSA)**

Die Verordnung über digitale Dienste (Digital Services Act, DSA) ersetzt die Richtlinie über den elektronischen Geschäftsverkehr. Ziel ist der bessere Schutz der Nutzerinnen und Nutzer und ihrer Grundrechte im Internet, die Transparenz und die Rechenschaftspflicht von Onlineplattformen sowie der einheitliche Rechtsrahmen in der EU. Plattformen werden verpflichtet, Massnahmen zur Bekämpfung illegaler Onlineinhalte vorzunehmen. Auch sollen die Plattformen Nutzenden Möglichkeiten zur Meldung illegaler Inhalte zur Verfügung stellen.

Die Regelungen gelten auch für Anbieter, die ausserhalb der Union niedergelassenen sind, aber ihre Dienste in der Union anbieten. Das Gesetz könnte auch Schweizer Intermediäre betreffen, wenn sie ihre Dienste in der EU bzw. im EWR anbieten oder darauf ausgerichtet sind. Um die freie Meinungsäusserung zu wahren, betreffen die Massnahmen nur illegale Inhalte, die im EU-Recht oder dem Recht der Mitgliedstaaten bereits als solches geregelt sind. Es werden keine neuen Straftatbestände eingeführt.

Sehr grosse Onlineplattformen (mind. 45 Millionen Nutzende) unterstehen der Regulierung der Kommission; Plattformen mit weniger Nutzenden der Aufsicht der Mitgliedstaaten, in welchen sie niedergelassen sind. Sehr grosse Onlineplattformen und Suchmaschinen werden verpflichtet, einen Missbrauch ihrer Systeme zu verhindern. Es besteht eine Pflicht zur Ergreifung risikobasierter Massnahmen und einer Beaufsichtigung in Form einer unabhängigen Prüfung. Zu mindern sind Risiken einer Desinformation, der Wahlmanipulation, der Cybergewalt gegen Frauen und jugendgefährdender Inhalte. Weiter wird ein Krisenreaktionsmechanismus vorgesehen, im Falle einer ernsten Bedrohung für die öffentliche Gesundheit und Sicherheit (Pandemie/Krieg). Gezielte Werbung aufgrund von Profiling von Kindern oder von besonders schützenswerten per-

---

[134] U.S. Kongress, DEEP FAKES Accountability Act, H.R.2395 — 117th Congress (2021–2022).

[135] Nicht eingegangen wird auf die unionsrechtlichen Regulierungen betreffend: Urheberrechtsschutz, Richtlinie über audiovisuelle Mediendienste, Datenschutzgrundverordnung, Grundrechtecharta.

sonenbezogenen Daten ist verboten. Die Verwendung von Dark Patterns an der Schnittstelle von Onlineplattformen wird ebenfalls verboten. Nutzerinnen und Nutzer haben das Recht, sich bei der Plattform zu beschweren, eine aussergerichtliche Streitbeilegung zu verlangen oder an die nationale Behörde zu gelangen und allenfalls Schadenersatz zu fordern. Sehr grosse Plattformen und sehr grosse Suchmaschinen, welche direkt der Aufsichts- und Durchsetzungsbefugnis der Kommission unterstehen, haben in den schwerwiegendsten Fällen Geldbussen bis zur Höhe von 6 % ihres gesamten Jahresumsatzes zu zahlen.

Kritik an der Verordnung bezieht sich unter anderem auf die Fehleranfälligkeit der Uploadfilter und die rechtmässige Entfernung von nicht illegalen Inhalten (Breyer 2023).

**Vorschlag für eine Verordnung über künstliche Intelligenz**[136]

Der Vorschlag der Europäischen Kommission für eine Verordnung über künstliche Intelligenz definiert Deepfakes als KI-Systeme, die Bild-, Ton- oder Videoinhalte erzeugen oder manipulieren, welche wirklichen Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen ähneln und echt oder wahrhaftig wirken. Der Vorschlag sieht für Deepfakes eine Transparenzpflicht vor. Seitens der Person, die ein Deepfake erschafft, ist gegenüber Nutzenden die künstliche Erzeugung oder Manipulation der Inhalte offenzulegen.[137] Eine Ausnahme besteht für Anwendungen, welche unter das Recht der freien Meinungsäusserung, der Kunstfreiheit oder der Wissenschaftsfreiheit fallen; die geeigneten Schutzvorkehrungen für die Rechte und Freiheiten Dritter vorausgesetzt.[138] Von der Transparenzpflicht befreit sind gesetzlich vorgesehene Verwendungen von KI-Systemen zur Aufdeckung von Deepfakes. Diese gelten als mit erweiterten Pflichten verbundene Hochrisiko-KI-Systeme.[139] Der Vorschlag sieht vor, dass gewisse, besonders schädliche KI-Systeme verboten werden. Deepfakes, welche zur unterschwelligen Beeinflussung des Bewusstseins einer Person eingesetzt werden oder ihr oder anderen physischen oder psychischen

[136] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union, Brüssel, den 21.4.2021, COM(2021)206 final, 2021/0106 (COD)

[137] Art. 52 Abs. 3 des Vorschlags (COM(2021)206 final).

[138] Art. 52 Abs. 3 des Vorschlags (COM(2021)206 final).

[139] Anhang III Ziff. 6 lit. c Anhänge des Vorschlags (COM(2021)206 final).

Schaden zufügen oder zufügen könnten, würden von diesem Anwendungsverbot des Vorschlags erfasst.[140]

Der Vorstoss steht jedoch auch in der Kritik. Insbesondere wird bemängelt, dass er sich ausschliesslich an die Nutzenden eines KI-Systems, nicht aber an die Hersteller und Anbieter dessen richtet. Im Hinblick auf das Herstellen und Anbieten von Deepfake-Produktionssoftware, die auf eine desinformative Nutzung abstellt, indem bewusst keine automatische Kennzeichnung erfolgt, entstünde so eine Regulierungslücke, weil Anbieter und Hersteller solcher Software nicht zur Rechenschaft gezogen werden könnten. Unklarheit herrscht auch darüber, wie die zur Rechtsdurchsetzung vorgesehenen Aufsichtsbehörden nicht gekennzeichnete Deepfakes, die von Nutzenden verbreitet wurden, erkennen können sollen. Auch bleibt der Einsatz von Detektionssoftware seitens der Plattformbetreiber ungeregelt (Karaboga 2023: 214 f.). Unklar ist ferner, wie die vom Vorschlag vorgesehene Offenlegung einer Manipulation umgesetzt werden soll und ob Hersteller der Technologie bei den Transparenzpflichten einzubinden sind: Die Anwendbarkeit einer Bestrafung nach Art. 71 des Vorschlags bei einem allfälligen Transparenzverstoss ist fraglich.[141]

Im Dezember 2023 haben sich Parlament, Kommission und Rat in den Trilog-Verhandlungen auf ein Kompromissdokument für den sog. AI Act geeinigt. Der Rechtsakt ist im Zeitpunkt der Abgabe des Manuskripts dieser Studie noch nicht von Parlament und Rat formell angenommen worden.

**Vorschlag für eine Verordnung über die Transparenz und das Targeting politischer Werbung (Europäische Kommission 2021a)**

Der Vorschlag für eine Verordnung über Transparenz und politisches Targeting soll unter anderem Desinformation und Manipulation sowie Eingriffe in Wahlen regulieren. Politische Werbung, welche auf Bearbeitung von personenbezogenen Daten mittels Targetings oder Amplifikation beruht, soll verboten werden. Es werden harmonisierte Transparenzpflichten vorgeschrieben, Transparenzlabels eingeführt und politische Werbung muss eindeutig gekennzeichnet werden. Targeting und Amplifizierung, also die Ausrichtung politischer Anzeigen auf bestimmte Personen oder Personengruppen, werden ebenfalls verboten, wenn diese auf sensiblen personenbezogenen Daten beruhen (Europäische Kommission 2021a).

---

[140] Art. 5; Art. 8 ff. des Vorschlags (COM(2021)206 final).

[141] Art. 52 Abs. 38 des Vorschlags (COM(2021)206 final).

**Verhaltenskodex zur Bekämpfung von Desinformation**

Der Verhaltenskodex betreffend Desinformation wurde in seiner überarbeiteten, verstärkten Version am 16. Juni 2022 unterschrieben. Unter den 34 unterzeichnenden Unternehmen befinden sich zum Beispiel Adobe, Google, Meta, Microsoft, TikTok und Twitter (EC 2022a). Der Kodex soll die Leitlinien zur Desinformation der EU-Kommission umsetzen. Unter dem DSA sollte der Kodex in Zukunft als Risikominderungsmassnahme in Form der Co-Regulierung anerkannt werden.

Der Kodex enthält 44 Verpflichtungen und 128 spezifische Massnahmen. Die Unterzeichner verpflichten sich unter anderem zur Demonetisierung und Limitierung der Desinformation, zur Gewährleistung der Transparenz politischer Werbung, zur verbesserten Zusammenarbeit mit Faktenprüfern in allen EU-Sprachen und zur Ermächtigung der Nutzer. Er sieht jedoch je nach Dienstleistung, Grösse und Wirkung der Plattformen unterschiedliche Verpflichtungen vor. Sehr grossen Onlineplattformen wird eine zentrale Rolle bei der Förderung der öffentlichen Debatte zugeordnet. Sie haben deshalb im Vergleich zu den kleineren Plattformen deutlich mehr Verpflichtungen wahrzunehmen. Gemäss DSA wäre zum Beispiel alle sechs Monate ein Bericht zur Umsetzung des Kodexes vorzulegen (EC 2022b). Im Gegensatz dazu sind kleinste Plattformen von der Pflicht, Beschwerde-, Rechtsbehelfsmechanismen und aussergerichtliche Streitbeilegungsverfahren einzurichten, befreit.

Beim Kodex handelt es sich um ein Instrument der branchenspezifischen Selbstregulierung. Die Kommission überwacht zwar die Fortschritte bei der Umsetzung des Kodex, kann allerdings keine Sanktionen bei Fehlverhalten aussprechen.

#### 4.5.2.3. Deutschland

Das deutsche Netzwerkdurchsetzungsgesetz ist seit dem 1. Oktober 2017 in Kraft. Das Netzwerkdurchsetzungsgesetz soll Hasskriminalität, strafbare Falschnachrichten und andere strafbare Inhalte im Internet bekämpfen. Es enthält gesetzliche Compliance-Regeln für soziale Netzwerke und regelt auch, wie Beschwerdeverfahren zu laufen haben. Die Anbieter von grossen sozialen Netzwerken haben die Pflicht, ein leicht erkennbares, unmittelbar erreichbares und ständig verfügbares Verfahren zur Übermittlung von Beschwerden anzubieten. Offensichtlich strafbare Inhalte sind innerhalb 24 Stunden nach Eingang der Beschwerde durch die Plattformen zu löschen oder zu sperren. Strafbare Inhalte sind grundsätzlich innert sieben Tagen zu löschen. Weiter sind die Anbieter von

grossen sozialen Netzwerken verpflichtet, halbjährlich einen öffentlich zugänglichen Bericht über die Anzahl der Beschwerden sowie über die Entscheidpraxis zu veröffentlichen. Der Bericht muss die Bemühungen des Intermediärs, kriminelle Aktivitäten auf seiner Plattform zu unterbinden, aufzeigen. Mechanismen zur Einreichung von Beschwerden sind zu beschreiben und die Entscheidungskriterien darzustellen.

Für eine wirksame Rechtsdurchsetzung sind die Unternehmen – unabhängig von ihrem Sitz – verpflichtet, für die Zustellung der Bussgeldverfahren in zivilrechtlichen Verfahren sowie für das Auskunftsersuchen der Strafverfolgungsbehörde einen inländischen Zustellungsbevollmächtigten zu benennen. Wird ein Beschwerdeverfahren nicht wirksam umgesetzt, kann die verantwortliche Person mit einer Geldbusse von bis zu 5 Millionen Euro belangt werden. Für eine Ordnungswidrigkeit des Unternehmens kann die Geldbusse bis zu 50 Millionen Euro betragen. Die Unternehmen sind zudem verpflichtet, besonders schwere Straftaten beim Bundeskriminalamt zu melden sowie zur Verfügung stehende Daten des Urhebers (letzte Log-In IP) zu übermitteln. Nutzer, deren gemeldeter Inhalt gelöscht werden sollte, können ein Gegenvorstellungsverfahren durchführen. Die Unternehmen müssen ihre Entscheidung überprüfen und begründen. Auch Personen, deren Meldung nicht gelöscht wurde, können eine Überprüfung anfordern. Streitigkeiten zwischen Nutzenden und sozialen Netzwerken sollen aussergerichtlich bei Schlichtungsstellen beigelegt werden. Plattformen können über das Gericht zur Datenherausgabe der benötigten Informationen über den Verfasser verpflichtet werden.

#### 4.5.2.4. Selbstregulierung der Plattformen

Einige Plattformen sehen auch freiwillige Massnahmen vor, um der Verbreitung von Des- und Misinformation[142] entgegenzuwirken und strafbaren Inhalt von ihren Plattformen fernzuhalten.

Zum Beispiel reguliert der *Facebook*-Gemeinschaftsstandard verschiedene Bereiche: Gewalt und kriminelles Verhalten, Sicherheit, anstössige Inhalte, Integrität und Authentizität und Wahrung des geistigen Eigentums. Gemäss dem Standard zu Integrität und Authentizität soll die Verbreitung von Fehlinformationen

[142] Die Begriffsverwendung der Plattformen ist nicht durchgehend konsistent mit den Definitionen dieses Berichts. Die Plattformen verzichten teilweise auf die Verwendung der einschlägigen Begriffe (Reddit), verwenden den Begriff Fehlinformation sowohl für Des- als auch Misinformation (TikTok) oder geben keine klare Definition (Facebook).

verringert oder ein Umfeld geschaffen werden, welches den produktiven Dialog fördert. Manipulierte Videos sollen entfernt werden, wenn sie so bearbeitet wurden, dass für eine Durchschnittsperson die Bearbeitung nicht mehr erkenntlich ist. Falschmeldungen, deren Inhalt wahrscheinlich direkt zur Beeinträchtigung des Funktionierens politischer Prozesse beiträgt, sowie stark betrügerische und manipulierte Medien werden gelöscht. Für die Einschätzung einer Fehlinformation arbeitet Facebook laut eigenen Angaben mit unabhängigen Experten zusammen (Meta 2023).

Auch *TikTok* kennt Community-Richtlinien, welche illegale Inhalte verbieten. Gemäss dieser werden Fehlinformationen entfernt, die das öffentliche Vertrauen in zivilgesellschaftliche Institutionen, Regierungen oder Wahlen untergraben. Weiter beinhalten die Community-Richtlinien ein Verbot digitaler Fälschungen (künstlich hergestellte oder manipulierte Medieninhalte), welche Nutzende in die Irre führen könnten, indem wahre Umstände einer Situation verzerrt werden oder erheblicher Schaden für die dargestellte Person, andere Personen oder die Gesellschaft entsteht (TikTok 2023).

Die Community-Richtlinien von YouTube verbieten Fehlinformationen, welche schädliche Folgen haben könnten, etwa indem sie demokratische Prozesse behindern. Verboten sind zudem technisch manipulierte Inhalte, welche Nutzende in die Irre führen oder ein ernsthaftes Risiko für Schäden darstellen. Technisch manipulierte Videos sind zulässig, wenn sie in einen entsprechenden Kontext eingebettet sind und etwa pädagogischen, dokumentarischen, wissenschaftlichen oder künstlerischen Zwecken dienen (Google Inc. 2023).

Die Richtlinien von *Reddit* sind bedeutend kürzer gehalten (Reddit 2023). Manipulationen von Inhalten und Wahlen sind grundsätzlich verboten, ebenso die Veröffentlichung intimer oder sexuell expliziter Inhalte ohne Einwilligung der betroffenen Personen und die Verwendung fremder Identitäten. Deepfakes in Form von Satire und Parodie sind je nach Einbettung in den Kontext erlaubt (Reddit 2020).

Die Plattformen ermöglichen es Nutzenden, falsche Nachrichten zu melden, welche dann in einem internen Verfahren überprüft und allenfalls entfernt werden. Gravierende Verstösse gegen die Community-Richtlinien können zu einer Kontosperrung des Nutzenden führen. Verschiedene Plattformen kennen ein Trusted-Flaggers-System, deren Antrag privilegiert behandelt wird. In der Schweiz kommt dem Bundesamt für Polizei (fedpol) eine solche Stellung bei YouTube zu. Der Bundesrat befürwortet eine Ausdehnung dieser Tätigkeit auf andere Social-Media-Plattformen (UVEK u.a. 2019: 33; Bundesrat 2017: 2).

Die Selbstregulierung von Intermediären wird kritisch gesehen, da befürchtet wird, dass Plattformen im Zweifelsfall zu viel Inhalt sperren (sog. «Overblocking») und es dadurch zu einer starken Beeinträchtigung von Kommunikationsgrundrechten kommen könnte (siehe Tschannen 2021: 400 N 1051). Nur offensichtlich rechtswidrige Inhalte sollen von Plattformen gelöscht werden. In umgekehrter Richtung wird aber auch befürchtet, dass Plattformen zu viel Zurückhaltung bei der Löschung von Inhalten an den Tag legen, weil sie keine nennenswerten Konsequenzen zu befürchten haben. Beide Kritikpunkte deuten auf eine Notwendigkeit staatlicher Vorgaben für die Plattformen hin. Weitere Punkte, die bei einer Selbstregulierung bemängelt werden, sind die folgenden: fehlende demokratische Mitwirkung, fehlende gerichtliche Kontrolle, Vernachlässigung öffentlicher Interessen, Gefahr einseitiger Machtausübung und kartellistischer Absprachen, Rechtsschutzdefizite, mangelnde Transparenz, Tendenz zu komplizierten und unübersichtlichen Regelungen, Gefahr der mangelnden Wirksamkeit und Effizienz (Marti 2000). Insgesamt als Vorteil zu werten ist das Vorhandensein des nötigen Fachwissens bei den Plattformen selbst, die Flexibilität der Regulierung, die internationale Ausrichtung, bessere Akzeptanz und Wirksamkeit sowie die Entlastung des Staates bezüglich der Regulierung (ebd.: 561 ff.).

Bei der Selbstorganisation der Intermediäre ist es nicht ersichtlich, welche Inhalte gelöscht werden. Es wird befürchtet, dass sie zu stark in Inhalte eingreifen, um allfälligen negativen Reaktionen vorzubeugen oder rechtliche Probleme zu vermeiden. Es bestehe die Gefahr, dass die angewandten Kriterien diskriminierend wirken könnten. Bei diesen internen Entscheidungsprozessen fehlt es zudem an verfahrensrechtlichen Minimalgarantien. Aus diesem Grund werden in der Literatur zusätzliche Regulierungen nach europäischem Vorbild gefordert, wie etwa Vorschriften betreffend Take-down-Prozeduren, Beschwerdemöglichkeiten oder das Einrichten von unabhängigen Überwachungsstellen (van der Sloot/Wagensveld 2022: 12; van Huijstee u.a. 2021: 62).

Aufgrund mangelnder Kapazitäten kann nur ein kleiner Teil der im Internet kursierenden Informationen durch Faktenchecker überprüft werden (Doublet 2019: 19).

Der Sonderberichterstatter der Vereinten Nationen für Meinungs- und Ausdrucksfreiheit empfiehlt, klare Regelungen bezüglich der Begrenzung des Contents, welcher die gesetzlichen Vorschriften überschreitet, vorzusehen. Zu Orientierung sollten objektive Kriterien und keine ideologischen oder politischen Ziele zugrunde liegen. Weiter sollten sie für einen einfachen Zugang zu Nutzungsbedingungen sorgen sowie klar Angaben zu deren Umsetzung veröffentlichen.

Nutzerinnen und Nutzer, deren Content gesperrt wird, sollten sofort informiert werden und die Möglichkeit haben, sich zu wehren (OSCE 2017).

Gleichzeitig besteht jedoch die Gefahr einer Überregulierung, wenn etwa Uploadfilter vorgeschrieben würden. Der Sonderberichterstatter der Vereinten Nationen rät von einer Verantwortlichkeit der Intermediäre ab, ausser im Falle eines aktiven Moderierens des Inhalts oder bei Verweigerung der Umsetzung eines Gerichtsurteils (van Huijstee u.a. 2021: 62; BAKOM 2011: 63; OSCE 2017).

# 5. Deepfakes im Journalismus

*Patric Raemy, Manuel Puppis & Gwendolyn Gurr*

Medien und Öffentlichkeit befinden sich in einem digitalen Strukturwandel (Habermas 2022; siehe auch Kapitel 6.1.1): Onlineplattformen wie soziale Netzwerke (bspw. Facebook, Instagram oder TikTok) oder Video-Sharing-Dienste (bspw. YouTube) spielen genauso wie Messengerdienste (bspw. WhatsApp oder Telegram) für die Information und Kommunikation der Bevölkerung heute eine zentrale Rolle (Newman u.a. 2022). Zudem verändern algorithmische Systeme und «schwache» künstliche Intelligenz die Produktion, Distribution und Konsumption von Information (Jungherr/Schroeder 2023; König/Wenzelburger 2020; Lin/Lewis 2022). Entsprechend wird nicht nur in der Politik, sondern auch in der Wissenschaft intensiv über die Chancen und Risiken dieser Veränderungen für demokratische Gesellschaften diskutiert.

Dieser Öffentlichkeitswandel hat potenziell nicht nur Implikationen für politische Prozesse in demokratischen Systemen (siehe Kapitel 6.1.2), sondern auch an Medienorganisationen gehen diese Entwicklungen nicht spurlos vorbei. Veränderte Nutzungsmuster und die damit zusammenhängende Verschiebung von Werbeeinahmen von Medien zu Plattformen haben massive Auswirkungen für die Refinanzierung journalistischer Angebote (Puppis u.a. 2017). Die Selektionslogik von Plattformen prägt die Nachrichtenauswahl in Redaktionen (Caplan/boyd 2018). Der Einsatz von Plattformen für Desinformationskampagnen erfordert eine teilweise aufwendige Verifikation (Schifferes u.a. 2014). Und die Möglichkeiten von KI verändern die redaktionelle Arbeit genauso wie die Personalisierungsmöglichkeiten für die Nutzenden (Borchardt 2022; Helberger 2019).

Auch synthetische Inhalte bieten für den Journalismus Chancen und Risiken. Eine besondere Form davon sind sog. Deepfakes, also mithilfe von KI-Techniken synthetisierte oder manipulierte Audio-, Bild- bzw. Videoinhalte, die authentisch wirken, es aber nicht sind (für Begriffsdefinitionen siehe Kapitel 1.3). Deepfakes gewinnen derzeit in der Kommunikationspraxis an Bedeutung und dürften immer schwerer zu erkennen sein (siehe Kapitel 2 für eine ausführliche Diskussion verschiedener Techniken zur Erstellung und Erkennung von Deepfakes). Sind Medienorganisationen und Medienschaffende dafür gewappnet, Deepfakes zu entlarven? Welche Möglichkeiten eröffnen Deepfakes für die redaktionelle Arbeit? Was passiert, wenn Journalistinnen und Journalisten mit Deepfakes belästigt werden? Und sehen sich Medien in der Lage, die Nutzenden für Deepfakes zu

sensibilisieren? Das folgende Kapitel ist Deepfakes im journalistischen Kontext gewidmet. Dabei geht es zuerst darum, den bisherigen Wissensstand im Sinne einer Status-quo-Analyse systematisch aufzuarbeiten und darzustellen, bevor die Resultate einer eigenen empirischen Studie zum Umgang mit Deepfakes in Schweizer Redaktionen präsentiert werden.

## 5.1. Theorie und Forschungsstand

### 5.1.1. Deepfakes und Journalismus

Journalismus wird eine wichtige Rolle für demokratische Systeme zugeschrieben. So wird vom Journalismus erwartet, dass er durch die Information über relevante Themen und Positionen zur Meinungs- und Willensbildung der Bürgerinnen und Bürger beiträgt, ein Forum für politische Debatten bietet und eine Kritik- und Kontrollfunktion gegenüber politischen Eliten wahrnimmt (Burkart 2021). Da normativer Anspruch und empirische Realität aber auch auseinanderklaffen können, schlägt Nielsen (2017: 1252) einen bescheideneren «demokratisch-realistischen Ansatz» vor: «The one thing I argue journalism just might do for democracy is this: provide people with relatively accurate, accessible, diverse, relevant, and timely independently produced information about public affairs.» Entsprechend gehört die auf professionellen Normen beruhende kritische Prüfung von Informationen, die in die Berichterstattung einfliessen, zu den Grundaufgaben journalistischer Arbeit. Diese Leistung des Journalismus kann auch dazu beitragen, die in der Politik wahrgenommenen Risiken von Deepfakes für die Demokratie zu minimieren (siehe Kapitel 6.2.2). Die zunehmende Bedeutung von Plattformen für die öffentliche Kommunikation hat die Überprüfung von Informationen allerdings nochmals erschwert. Es erstaunt deshalb nicht, dass Deepfakes in der Journalismusforschung bisher hauptsächlich im Kontext der Faktenprüfung thematisiert und mit Desinformation in Verbindung gebracht werden. Mit Desinformation wird die intentionale Verbreitung faktisch falscher Informationen aus politischen oder kommerziellen Gründen bezeichnet. Eine Spezialform davon sind sog. «Fake News», also Desinformation, die aussieht wie ein journalistischer Inhalt (Tandoc Jr u.a. 2019; Kalsnes u.a. 2021, siehe auch Kapitel 6).

Der digitale Strukturwandel der Öffentlichkeit und die dadurch erleichterte Verbreitung von Desinformation und «Fake News» hat in der Kommunikations- wie der Politikwissenschaft eine intensive Beschäftigung mit der Thematik ausgelöst (bspw. Guess u.a. 2019; Lazer u.a. 2018; Nelson/Taneja 2018; Tandoc Jr u.a.

2019; Tsfati u.a. 2020; Vu/Saldaña 2021). Diskutiert werden Auswirkungen auf die Demokratie und politische Prozesse, die Kommunikation politischer Akteure, Fragen der Governance sowie die Erkennung durch einzelne Nutzerinnen und Nutzer. Auch Redaktionen stellt die vereinfachte Möglichkeit zur Verbreitung faktisch falscher Informationen in der Öffentlichkeit vor grosse Herausforderungen (van Aelst u.a. 2017). Desinformation erfordere es, die Rolle von Journalismus in der Demokratie ganz grundsätzlich zu überdenken (Waisbord 2018). Auch wenn es an der behaupteten Zunahme von Desinformation Zweifel gibt und «Fake News» als Spezialform von Desinformation nicht zum Medienrepertoire der meisten Menschen gehören (Allen u.a. 2020; Williams 2023; Lecheler/Egelhofer 2022; Grinberg u.a. 2019): Der Journalismus steht vor der Herausforderung, Desinformation zu erkennen, nicht zuletzt, weil bei deren (unabsichtlicher) Weiterverbreitung auch die eigene Glaubwürdigkeit leidet. Insofern stellt Desinformation eine Gefahr für die Legitimität des Journalismus dar (Schapals/Bruns 2022), bietet aber auch die Chance, sich durch auf professionellen Normen beruhende Verifikationsprozesse von alternativen Quellen zu unterscheiden.

Das Forschungsinteresse in vielen Disziplinen liegt auf der Erkennung von Deepfakes (siehe Kapitel 2.6.2 und 2.8). Auch in der Journalismusforschung liegt der Fokus bezüglich Deepfakes hauptsächlich auf den Herausforderungen für deren Erkennung durch Redaktionen. Mit Deepfake-Technologien hat sich Desinformation von meist schriftlichen Artikeln und Kommentaren oder statischen Bildern hin zu synthetischen Video- und/oder Audioinhalten entwickelt, die schwer von unveränderten Medieninhalten zu unterscheiden sind: «Sophisticated, deep-learning techniques using forms of artificial intelligence to create deliberately distorted audio-visual material known as ‹deepfakes› are likely to intensify the issue of mis- and disinformation» (Schapals/Bruns 2022: 7).

Deepfakes können auf Plattformen schnell verbreitet werden (Godulla u.a. 2021; Sylvester 2021; Vizoso u.a. 2021). Auf den Journalismus kommt deshalb die Aufgabe zu, Videos in kürzester Zeit zu verifizieren, bevor diese von der Öffentlichkeit als real wahrgenommen werden (Sohrawardi u.a. 2020). Eine Verifizierung ist auch wichtig, um zu vermeiden, dass Medien selbst zur Weiterverbreitung von Deepfakes beitragen. Einmal veröffentlicht, wird es Redaktionen schwerfallen, ihre Berichterstattung zu korrigieren. Deepfakes können zu einem Kontrollverlust der Redaktion führen (Godulla u.a. 2021), wenn beispielsweise ein durch die Redaktion verbreiteter Deepfake über andere Kanäle (und mit dem Medium als glaubwürdige Quellenangabe) weiterverbreitet wird. Des Weiteren wird auf die Gefahr hingewiesen, dass durch eine erhöhte Thematisierung von bestimmten Deepfakes diese erst recht Aufmerksamkeit erlangen (Wahl-Jorgensen/Carlson 2021).

Doch auch indirekt haben Deepfakes Auswirkungen auf den Journalismus. Zum einen können Deepfakes über ihren Einfluss auf einzelne Nutzende zu einer Abnahme von Vertrauen in die Medien führen (Collins/Ebrahimi 2021; Godulla u.a. 2021; Gutsche 2019; Wahl-Jorgensen/Carlson 2021; Westerlund 2019). Zunächst erschweren Deepfakes das Unterscheiden richtiger und falscher Informationen (Yadlin-Segal/Oppenheim 2021), sorgen bei Nutzerinnen und Nutzern für Unsicherheit (Vizoso u.a. 2021) und erschweren die Qualitätseinschätzung von Nachrichten online (Godulla u.a. 2021; Vizoso u.a. 2021). Zum anderen können Deepfakes die öffentliche Meinung und politische Kampagnen beeinflussen (Godulla u.a. 2021) und von (politischen) Akteuren zur Verbreitung von Propaganda eingesetzt werden (Wahl-Jorgensen/Carlson 2021). Deepfakes wird deshalb das Potenzial zugesprochen, den Sinn für soziale Realität abzuschwächen (Yadlin-Segal/Oppenheim 2021) und die Authentizität des politischen Diskurses online einzuschränken (Vaccari/Chadwick 2020). Hinzu kommt die Gefahr, dass (para-)journalistische Angebote, die sich durch eine einseitige oder gezielt irreführende Berichterstattung auszeichnen, vermehrt Deepfakes einsetzen, um das Vertrauen in die Medien zu untergraben. Zum Beispiel zeigen Humprecht et al. (2023), dass populistische Akteure und deren (journalistische) Kommunikationskanäle oft Fehlinformationen verbreiten, um ihre Gegner zu diskreditieren, wobei ihr Publikum zugleich anfälliger dafür ist, Desinformation zu glauben. Diese Entwicklungen stellen für die Rolle des Journalismus in der Demokratie Risiken dar.

*Redaktionellen Arbeitsprozessen und Strategien* kommt bei der Erkennung von Deepfakes eine zentrale Rolle zu. Um die gesellschaftlichen Funktionen des Journalismus erbringen zu können, haben Redaktionen Strukturen ausgebildet, innerhalb derer Medienschaffende auf Grundlage bestimmter Regeln und Routinen (bspw. für die Auswahl von Nachrichten oder die Prüfung von Informationen) Inhalte produzieren (Altmeppen 2006; Shoemaker/Reese 2014). Journalistinnen und Journalisten sollten sich bei ihrer Arbeit an bestimmten professionellen Normen orientieren, um ihre Funktionen für die Demokratie zu erfüllen (Christians u.a. 2009). Die Umsetzung journalistischer Ideale gelingt allerdings bei Weitem nicht immer (Raemy u.a. 2021). So sind Journalistinnen und Journalisten in ihrem Arbeitsalltag oft mit Zielkonflikten zwischen journalistischen Normen und kommerziellen Zielen von Medienorganisationen konfrontiert (McChesney 2008; Raemy/Vos 2021). Zudem hat die Finanzierungskrise des Journalismus den Spardruck auf die Redaktionen massiv erhöht (Puppis u.a. 2017), was auch negative Auswirkungen auf die sorgfältige Prüfung von Informationen haben kann.

Nun stellt sich die Frage, inwiefern diese Arbeitsprozesse und Routinen in Redaktionen angepasst werden müssen, um auch Deepfakes erkennen zu können

und nicht unbeabsichtigt zu deren Weiterverbreitung beizutragen. Zunächst können Medienunternehmen durch verschiedene Kooperationen die Risiken von Deepfakes verringern. Kooperationen mit Forschung, Plattformunternehmen und anderen Medienorganisationen werden als Strategie diskutiert (Vizoso u.a. 2021). Dabei können gemeinsame Datenbanken (Sylvester 2021) und Blogs (Vizoso u.a. 2021) hilfreich sein. Diese Strategien gehen mit der allgemeinen Empfehlung einher, vermehrt auf Machine Learning und Technologie zu setzen (Gregory 2019; Leibowicz u.a. 2020, July 13; Wahl-Jorgensen/Carlson 2021). Konkret werden Algorithmen (Collins/Ebrahimi 2021), Computational Methods (Vizoso u.a. 2021), forensische Techniken und digitale Wasserzeichen (Westerlund 2019) sowie Blockchain-basierte Deepfake-Scanner (Sylvester 2021) diskutiert (vgl. auch die Ausführungen in Kapitel 2). Godulla u.a. (2021) tragen in ihrer Literaturanalyse diverse detaillierte Schritte zur Erkennung von Deepfakes zusammen, z.B. auf Schlüsselstellen im Video zu fokussieren oder Gesichtsausdrücke und Mimik zu analysieren. Allerdings besteht das Risiko, dass Instrumente zur Erkennung von Deepfakes nicht akkurat funktionieren und nicht alle Fälschungen entdecken (Sohrawardi u.a. 2020; siehe auch Kapitel 2.7). Zudem können nicht manipulierte Inhalte als falsch erkannt werden (Lewis u.a. 2022), was als eine Einschränkung der Meinungsfreiheit gelten kann (Diakopoulos/Johnson 2021; Sohrawardi u.a. 2020).

Als ebenso bedeutend wie der Einsatz technischer Hilfsmittel gilt die Sensibilisierung und Schulung der Medienschaffenden im Umgang mit Deepfakes (Marconi/Daldrup 2018), wie z.B. das Arbeiten an Trainingsdatensätzen zur versuchsweisen Herstellung von und Sensibilisierung für Deepfakes (Gregory 2019; Malik u.a. 2022; Lyu 2020). Hierbei kommt auch journalistischen Ausbildungsorganisationen eine wichtige Rolle zu.

Für den Journalismus ist es also zentral, Deepfakes zu erkennen. Zusätzlich müssen diese auch richtig eingeordnet werden. Nicht bei jedem Deepfake handelt es sich um eine Form der Desinformation mit Täuschungsabsicht, sondern Deepfakes können auch zur Unterhaltung, als Form der Satire oder mit künstlerischem Anspruch erstellt werden. Zudem müssen Redaktionen jeweils prüfen, ob zivil- oder strafrechtliche Vorschriften gegen eine Weiterverbreitung sprechen resp. in welchem Rahmen eine Berichterstattung möglich ist. Hierbei ist insbesondere an Persönlichkeitsverletzungen, Identitätsmissbrauch oder pornografische Darstellungen zu denken (siehe Kapitel 4.2). Für den Rundfunk gilt zudem das Sachgerechtigkeitsgebot (siehe Kapitel 4.4). Innerhalb dieses rechtlichen Rahmens sind Medien in ihrer Berichterstattung über Deepfakes inkl. deren Weiterverbreitung frei. Umso wichtiger ist die Einhaltung professioneller jour-

nalistischer Normen, um einen verantwortungsvollen Umgang mit Deepfakes zu gewährleisten, der auch den potenziellen Auswirkungen auf die Meinungsbildung in der Bevölkerung gerecht wird.

Schliesslich können Medienunternehmen und Medienschaffende die Risiken auch einzudämmen versuchen, indem sie einen Beitrag dazu leisten, ihr Publikum für Deepfakes zu sensibilisieren (Diakopoulos/Johnson 2021; Godulla u.a. 2021; Wahl-Jorgensen/Carlson 2021). Wie im Rahmen dieser Studie gezeigt wurde, hat ein Grossteil der Bevölkerung bisher nur wenig Wissen über resp. Erfahrung mit Deepfakes (siehe Kapitel 3): Nur rund die Hälfte der Bevölkerung kennt den Begriff Deepfake und hat nach eigenen Angaben schon Deepfakes gesehen. Zudem herrscht bei der Beurteilung von Videos grosse Unsicherheit. Vizoso u.a. (2021) schlagen etwa vor, Beispielsmaterial für Nutzende zur Verfügung zu stellen, um sie auf Deepfakes zu sensibilisieren. Auch das Veröffentlichen von falschen Videos zusammen mit korrekten Videos kann hilfreich sein (Westerlund 2019).

Deepfakes bringen aber nicht nur Risiken für die öffentliche Meinungsbildung und das Vertrauen in die Medien mit sich, sondern sie können auch *Medienschaffende selbst betreffen*. So können Deepfakes dazu benutzt werden, um Journalistinnen und Journalisten anzugreifen und kritische Berichterstattung zu verhindern (Posetti 2018a; Posetti 2018b). Insbesondere Journalistinnen sehen sich in der Öffentlichkeit mit besonders heftigen Angriffen konfrontiert (Posetti u.a. 2021), was auch eine akute Bedrohung für die Medienfreiheit darstellt. Mit Deepfakes steht für Angriffe ein weiteres Mittel zur Verfügung: Journalistinnen und Journalisten laufen Gefahr, durch Deepfakes in Misskredit gebracht zu werden (Ayyub 2018; Compton 2021; Marconi/Daldrup 2018; Razek 2018; Suvorova 2022). So ist es denkbar, sie in pornografische Deepfakes zu integrieren, mit falschen Interviewpartnern zusammenzubringen, als falsche Interviewerinnen in einem Deepfake zu erstellen oder ihr Handeln in Interviews zu manipulieren. Ebenso können Deepfakes genutzt werden, um vertrauenswürdige Persönlichkeiten wie Journalistinnen für Werbezwecke zu missbrauchen. Solche Angriffe auf Medienschaffende verletzen auch Persönlichkeitsrechte (siehe Kapitel 4.2.1).

Doch Deepfakes resp. die dahinterstehende Technologie kann auch *Chancen für den Journalismus* bieten. Duffy/Ang (2019) halten es nicht nur für möglich, dass die Nutzenden das Vertrauen in Nachrichtenjournalismus verlieren bzw. sich von Nachrichten insgesamt abwenden könnten, sondern auch, dass sich das Publikum gerade wegen der Unsicherheit im Internet wieder vermehrt traditionell glaubwürdigen Medien zuwendet. Auch Wahl-Jorgensen/Carlson (2021) führen an, dass ein erfolgreicher Umgang des Journalismus mit Deepfakes eine

Abgrenzung zu sozialen Netzwerken ermöglicht, wo Deepfakes ungefiltert verbreitet werden können.

Jenseits der Problematik für Redaktionen, Deepfakes zu erkennen, können Deepfake-Technologien im Journalismus auch für Unterhaltungszwecke und kreative Anwendungen genutzt werden (siehe auch Kapitel 7.2) und damit die Personalisierung, Visualisierung und Immersion von Nachrichten verbessern (Collins/Ebrahimi 2021; Gamage u.a. 2022a; Hellyer 2022; Wahl-Jorgensen/Carlson 2021). Ebenso können verstorbene Personen durch die Technologie wieder sprechend in Videos präsentiert werden (Willi Kägi 2022; Suvorova 2022; Sylvester 2021).[143] Deepfakes vereinfachen auch die Produktion von Inhalten in mehreren Sprachen (Diakopoulos/Johnson 2021) oder die Übersetzung von Inhalten in Gebärdensprache mit Avataren (siehe das Beispiel von SwissTXT, o.J.). Entsprechend ist zu betonen, dass das Thema Deepfakes auch im Journalismus nicht nur im Kontext von Desinformation zu sehen ist. Allerdings wirft die Automatisierung und Veränderung journalistischer Inhalte wichtige ethische und urheberrechtliche Fragen auf (Montal/Reich 2017).

Und schliesslich bietet künstliche Intelligenz als Haupttechnologie hinter Deepfakes auch ein Potenzial zur Erkennung manipulierter Inhalte (bspw. Huang u.a. 2022; Jarrahi/Safari 2022; Peng/Xintong 2022; Rohera u.a. 2022). So wäre es denkbar, dass in Medienunternehmen entwickelte Tools vermarktet werden können.

### 5.1.2. Forschungsstand: Was wir bisher wissen

Zum Thema Deepfakes und Journalismus gibt bisher nur sehr wenige empirische Studien, wobei meist der Diskurs in den (sozialen) Medien (Gamage u.a. 2022b; Westerlund 2019; Wahl-Jorgensen/Carlson 2021; Yadlin-Segal/Oppenheim 2021) oder die Erkennung durch resp. Auswirkungen auf die Nutzenden im Vordergrund stehen (Ahmed 2021a; Dobber u.a. 2021; Lewis u.a. 2022; Vaccari/Chadwick 2020; siehe Kapitel 3). Aufgrund der zentralen Rolle des Journalismus bei der (Verhinderung der) Weiterverbreitung von Deepfakes sind aber auch der Umgang mit Deepfakes in Redaktionen und Strategien zu deren Erkennung zentral. Dennoch ist nur wenig darüber bekannt, wie Journalistinnen und Journalisten die Auswirkungen von Deepfakes auf die journalistische Arbeit wahrnehmen und wie Redaktionen darauf reagieren.

[143] Siehe auch die TA-SWISS-Studie «Tod im digitalen Zeitalter»: Strub u.a. (2024).

Vizoso u.a. (2021) haben mit einer Dokumentenanalyse untersucht, wie drei grosse Nachrichtenorganisationen (Wall Street Journal, Washington Post, Reuters) und Plattformunternehmen (Google, Facebook, Twitter) mit Deepfakes umgehen. Die untersuchten Medien setzen auf die Weiterbildung der Journalistinnen und Journalisten zur Erkennung von Deepfakes (insb. auch mittels technischer Instrumente). Dabei wird auf verschiedene Massnahmen wie die Untersuchung von Quellen, die Suche nach älteren Versionen oder die Untersuchung von Material mit Foto- und Videobearbeitungsprogrammen gesetzt. Hierfür wird auch mit Wissenschaft und Plattformunternehmen kooperiert. Plattformen hingegen investieren in Forschungsprojekte, die der Verbesserung forensischer Werkzeuge dienen.

Sohrawardi u.a. (2020) nutzten Interviews mit für die Verifizierung von Informationen zuständigen Personen bei verschiedenen Medien, um mehr über den derzeitigen Umgang mit Deepfakes und die Anforderungen an Erkennungssoftware zu erfahren. Dabei zeigte sich, dass bislang nur ein Tool zur Videoverifizierung (InVid) eingesetzt wird, ansonsten auf den Abgleich von Metadaten und eigene Beobachtung vertraut wird. Mit Blick auf Erkennungssoftware sind den Journalistinnen und Journalisten insbesondere Fehlerfreiheit und Erklärbarkeit der Resultate wichtig.

Schon vor dem Aufkommen von Deepfakes haben Brandtzaeg u.a. (2016) Interviews mit Journalistinnen und Journalisten in verschiedenen europäischen Ländern zur Prüfung von Informationen auf Onlineplattformen geführt. Dabei wurde auch die *Verifizierung von Bildern und Videos* thematisiert. Damals wurde vor allem versucht, ergänzende Informationen über die Person, welche ein Video bereitstellt, zu beschaffen oder deren Profil zu prüfen. Die befragten Journalistinnen und Journalisten bekundeten bei der Verifikation vor allem Mühe mit der Flut an Information und der zunehmenden technischen Komplexität. Nur teilweise wurden Onlinetools (wie bspw. ein Abgleich von Bildern und Videos mit Google Street View) eingesetzt. Zur Erkennung von Deepfakes wird das aber nicht ausreichen.

Wie bereits erwähnt, werden Deepfakes in der Journalismusforschung zumeist im Kontext der Erkennung von Desinformation diskutiert, weshalb auch ein Blick auf Erkenntnisse über den Umgang von Redaktionen mit Desinformation und «Fake News» lohnt. Mehrere Studien haben mittels Interviews mit der Chefredaktion und Journalistinnen die Reaktion von Medienorganisationen auf die Delegitimierung von Journalismus durch Desinformation untersucht, so in Deutschland, Grossbritannien, Skandinavien, Rumänien, Australien und den USA. Medienorganisationen versuchen demnach

- die Bedeutung von Qualitätsjournalismus für die Demokratie und die Reputation des eigenen Mediums zu betonen (Jahng u.a. 2023; Kalsnes u.a. 2021),
- die Transparenz der Nachrichtenproduktion und Visibilität professioneller Normen und Prozesse zu erhöhen (Jahng u.a. 2023; Koliska/Assmann 2021; Schapals/Bruns 2022),
- die eigenen professionellen Standards zu verbessern, insbesondere durch eine intensivere Verifizierung von Informationen, wie beispielsweise durch Aus- und Weiterbildung im Bereich der Open-Source Intelligence (Koliska/Assmann 2021; Schapals/Bruns 2022),
- die Medienkompetenz (v.a. bei jungen Menschen) zu fördern (Dumitru 2021; Kalsnes u.a. 2021; Schapals/Bruns 2022) oder auch
- in einen Austausch mit der eigenen Community zu treten (Jahng u.a. 2023).

Zudem hat der US-amerikanische Autorenverband PEN America (2022) eine Umfrage unter Journalistinnen und Journalisten zu den Auswirkungen von «Fake News» in Auftrag gegeben. Diese zeigt, dass Medienschaffende mit sinkendem Vertrauen und mit Anfeindungen aus der Öffentlichkeit konfrontiert sind (was auch dazu führen kann, dass über ein Thema aus Angst vor Fake-News-Vorwürfen nicht berichtet wird), dass Journalisten Desinformation auch selbst immer wieder aufsitzen oder mit den notwendigen Verifikationsschritten überfordert sind, dass sich Arbeitsprozesse verändert haben (insb. ist mehr Zeit für die Prüfung von Informationen nötig) und dass stärker auf Transparenz der Nachrichtenproduktion geachtet wird. Gleichzeitig haben aber bei Weitem nicht alle Medienorganisationen tatsächlich Massnahmen zum Umgang mit Desinformation getroffen.

Dieser Überblick verdeutlicht, dass der Einfluss von Deepfake-Technologien sowie einzelner Deepfakes auf die redaktionelle Arbeit in Medienorganisationen und auf die Aus- und Weiterbildung von Medienschaffenden bisher nicht umfassend untersucht wurde. Zu den konkreten Massnahmen, die Redaktionen für den Umgang mit Deepfakes getroffen haben, liegen kaum Erkenntnisse vor. Studien mit Bezug zur Schweiz fehlen gänzlich. Ebenso finden sich keine Studien dazu, wie Medienorganisationen mit Situationen umgehen, in denen Journalistinnen und Journalisten selbst von Deepfakes betroffen sind. In diesem Kapitel werden daher die folgenden drei Forschungsfragen untersucht:

- **FF 4.1:** Welche Strategien werden im Journalismus zur Identifikation von Deepfakes angewendet und zu welchen konkreten Anpassungen von Arbeitsprozessen und Routinen führt dies in Redaktionen von Schweizer Medienorganisationen aktuell und in Zukunft?
- **FF 4.2:** Wie werden die Herausforderungen, welche Deepfakes für den Journalismus mit sich bringen, in der Ausbildung von Medienschaffenden aktuell und in Zukunft thematisiert?
- **FF 4.3:** Wie sind Medienorganisationen auf Fälle vorbereitet, in denen Journalistinnen und Journalisten selbst von Deepfakes betroffen sind?

## 5.2. Methodische Vorgehensweise

Zur Beantwortung der drei Fragestellungen kamen im empirischen Teil der Studie mehrere Methoden zum Einsatz.

Mit Blick auf den *Umgang mit Deepfakes in der journalistischen Ausbildung* (FF 4.2) wurden die wichtigsten unabhängigen Ausbildungsorganisationen in der Schweiz (AJM/Universität Neuenburg, CFJM, IAM/ZHAW, IMP/FHGR, MAZ, Corso di Giornalismo) auf der Basis von Dokumenten (Reh 1995) und Recherchegesprächen betrachtet. Die Ausbildungsorganisationen wurden im Juli 2022 angeschrieben, um einen Termin für ein Telefongespräch zu planen sowie mit der Bitte, Dokumente zu Kursen, in denen Deepfakes thematisiert werden, zuzusenden oder Hinweise zu geben, wo entsprechende Dokumente verfügbar sind. Ziel war es, Faktenwissen zusammenzutragen, weshalb eine Auswertung der Dokumente im Sinne eines «literal readings» ausreichend war (Mason 2018). Dabei wurde versucht, anhand der Informationen aus den Gesprächen und Dokumenten herauszufinden, (1) welche spezifischen Kurse und Lerneinheiten zu Deepfakes angeboten werden, (2) wie Deepfakes thematisiert werden und (3) welchen Stellenwert Deepfakes in der Ausbildung aktuell haben. Die Informationen wurden pro Ausbildungsorganisation zusammengefasst und dargestellt.

Der Schwerpunkt der empirischen Arbeit lag aber auf den Medienorganisationen selbst, also dem *Umgang von Redaktionen verschiedener Schweizer Medien mit Deepfakes* (FF 4.1), sowie Situationen, in denen *Journalistinnen und Journalisten selbst von Deepfakes betroffen* sind (FF 4.3). In der vorliegenden Untersuchung kam mit Experteninterviews eine besondere Form teilstandardisierter Interviews zum Einsatz. Als Expertinnen und Experten gelten Personen, die Angehörige einer Funktionselite innerhalb eines organisatorischen Kontextes

sind – also Personen, die durch ihr Wissen und ihre Position in einer Organisation Einfluss besitzen (Bogner/Menz 2002; Meuser/Nagel 1991). Im Vordergrund solcher Interviews steht die kommunikative Erschliessung und analytische Rekonstruktion von subjektiven Deutungen und Interpretationen. Aufgrund der beschränkten finanziellen Mittel konnten nur die grössten Medienorganisationen in der Deutschschweiz und der Suisse Romande berücksichtigt werden. Die Medienorganisationen wurden im September 2022 mit der Bitte angeschrieben, eine geeignete Person für ein Interview zu vermitteln, die über den Umgang ihrer Redaktionen mit Deepfakes Auskunft geben kann. Zugesagt haben insgesamt zehn Personen aus neun Redaktionen. Interviewt wurden Vertreterinnen und Vertreter der folgenden Medienorganisationen und Redaktionen[144] (siehe Tabelle 4):

**Tabelle 4:** Untersuchte Medienorganisationen und Redaktionen

| Organisation | Redaktion | Funktion der Person |
|---|---|---|
| AZ Medien | Watson | Führungsperson Redaktion |
| Fondation Aventinus | Heidi.news | Redaktion im Bereich Wissenschaft und Technologie |
| NZZ Mediengruppe | Neue Zürcher Zeitung | Open Source und Visual Investigations |
| Ringier | Blick | Führungsperson Recherche |
| SRG SSR | Person 1: SRF, Abteilung Recherche und Archive | Person1: Führungsperson Faktencheck |
| | Person 2: SRF, Abteilung Produktion und Technologie | Person 2: Produktmanagement Artificial Intelligence |
| | RTS, Actualité & Sports | Responsable coordination numérique Actu |
| TX Group | 20 Minuten | Wissensredaktion |
| | 20 minutes | Führungsperson Redaktion |
| | Tamedia Zentralredaktion Deutschschweiz | Führungsperson Videoredaktion |

[144] Ein grosses Medienunternehmen in der Deutschschweiz entschied sich gegen eine Teilnahme an der Studie.

Im Zentrum der Gespräche stand, wie die Interviewten Chancen und Risiken von Deepfakes und der dahinterstehenden Technologie im Journalismus wahrnehmen und wie ihre Redaktionen damit umgehen. Zur Durchführung der Gespräche wurde aus dem Stand der Forschung ein Leitfaden entwickelt, der die Forschungsfragen in Themenbereiche und Gesprächsfragen übersetzt (siehe Anhang A.2). Der Leitfaden enthielt Fragen zu den folgenden Themenbereichen:

- Einschätzung von Chancen und Risiken von Deepfakes;
- Prävention innerhalb der Redaktion resp. der Medienorganisation;
- Erkennung von Deepfakes und Verhinderung der Weiterverbreitung;
- Journalistinnen und Journalisten als Betroffene;
- Massnahmen nach der Veröffentlichung und Verbreitung durch Medien;
- Sensibilisierung der Nutzerinnen und Nutzer und Förderung von Medienkompetenz.

Die Gespräche dauerten ungefähr 60 Minuten. Zuerst wurde offen nach der Einschätzung von Chancen und Risiken von Deepfakes gefragt. Danach wurde jeweils eine Liste mit möglichen Chancen und Risiken gezeigt, welche die Interviewten einschätzen sollten. Die Liste wurde aus dem Stand der Forschung abgeleitet. Es folgte ein Themenblock zur Einschätzung der Prävention der Verbreitung von Deepfakes innerhalb der Redaktion bzw. der Medienorganisation, ein Themenblock zur Erkennung und Filterung von Deepfakes, ein Themenblock zu den Massnahmen, wenn Journalisten selbst von Deepfakes betroffen sind, ein Themenblock zu Interventions- und Reparationsmöglichkeiten, für den Fall, dass ein Medium tatsächlich unwillentlich Deepfakes verbreitet, und ein letzter Themenblock, der die Sensibilisierung und Förderung von Medienkompetenz beim Publikum thematisierte.

Die Interviews wurden zwischen Oktober 2022 und Januar 2023 durchgeführt. Alle Interviews wurden aufgezeichnet und anschliessend transkribiert.[145] Die Interviews wurden mittels einer induktiven qualitativen Inhaltsanalyse ausgewertet. Dabei werden Kategorien (siehe Anhang A.3) in einem Zusammenspiel aus induktiver und deduktiver Vorgehensweise entwickelt (Mayring 2010; Wagner/Schönhagen 2021; Puppis 2019). Mithilfe der Software MAXQDA wurden dem Textmaterial Kategorien zugeordnet und anschliessend wurde mittels eines thematischen Vergleichs eine Interpretation der Daten vorgenommen. Alle verwendeten Zitate wurden den Interviewten zur Autorisierung vorgelegt.

[145] Für die Transkription danken wir Sabrina Heiland und Hanna Wenger.

## 5.3. Resultate

Die Resultate zeigen, dass Deepfakes als Phänomen in der Journalismusausbildung und in den Redaktionen zwar wahrgenommen wird, zum Zeitpunkt der Erhebung aber nicht als dringendes Risiko thematisiert und priorisiert wurde. Zudem werden Deepfakes als ein Aspekt von Desinformation eingeschätzt. Im Umgang mit Deepfakes wird die Bedeutung der Einhaltung grundlegender journalistischer Standards betont. Dabei beziehen sich die Ausbildungsorganisationen wie auch die interviewten Personen auf die grundlegende Norm im Journalismus, Information so gut wie möglich nach ethischen und professionellen Standards zu recherchieren, zu validieren, aufzubereiten und zu veröffentlichen. Diese Norm und die berufliche Identität, eine verantwortungsvolle Informationsquelle für die Nutzenden zu sein, sollen sowohl in der Ausbildung wie auch in der redaktionellen Praxis internalisiert werden. Insgesamt zeigte sich, dass das Thema relativ neu ist und entsprechend noch nicht bei allen Redaktionen und Ausbildungsorganisationen eine prioritäre Rolle spielt.

### 5.3.1. Thematisierung der Herausforderungen durch Deepfakes in der Journalismusausbildung

Bisher befassen sich in der Ausbildung von Journalistinnen und Journalisten tätige Organisationen in der Schweiz nicht explizit mit Deepfakes, fördern aber im Rahmen ihrer Angebote wichtige journalistische Kompetenzen, welche im Umgang mit Deepfakes hilfreich sein können. Dazu gehören vor allem zwei Schwerpunkte:

- *Verantwortungsvolle Verifikation von Informationen:* Durch die Förderung von Medienethik und kritischem Denken sollen angehende Journalisten fähig sein, Information zu verifizieren. Im Zusammenhang mit manipulierten Fotos und Videos werden Deepfakes oft als Beispiel herangezogen. Durch die Sensibilisierung für manipulierte Information soll das Risiko minimiert werden, selbst zum Sender von Falschinformationen zu werden, aber auch zur unwissenden Empfängerin.
- *Deepfakes und die aktuellen technologischen Möglichkeiten verstehen:* Die Studierenden sollen Techniken kennen, um veränderte Fotos und gefälschte Videos zu erkennen. Sie sollen verstehen, wie solche Videos hergestellt werden. Sie sollen Fake-Fotos und Deepfake-Videos erkennen und selbst produzieren können, um diese Art von Technologie besser zu verstehen. Sie

sollen Prozesse verstehen, durch die gefälschte Bilder, Videos und Töne entstehen. Dabei sollen sie auch die Geschichte von Desinformation kennen (historischer Kontext).

Das Centre de formation au journalisme et aux médias (CFJM) in Lausanne ist bisher die einzige Ausbildungsorganisation, welche das Thema Deepfakes explizit in einem halbtägigen Kurs im Rahmen der Ausbildung von Praktikantinnen und Praktikanten im richtigen Umgang mit sozialen Netzwerken thematisiert. Zusätzlich bietet das CFJM seit 2021 einen eintägigen, öffentlich zugänglichen Kurs zum Thema Deepfakes an (CFJM 2022).

Bei allen anderen Ausbildungsorganisationen wurden Deepfakes zum Zeitpunkt der Erhebung als Beispiel für verwandte Themen verwendet. Das Institut für Multimedia-Production (IMP) der Fachhochschule Graubünden (FHGR) integriert das Thema bei anderen thematisch verwandten Modulen und Kursen wie zum Beispiel der Medienethik.

Von der Associazione Corso di giornalismo della Svizzera italiana werden mit Deepfakes verwandte Themen wie künstliche Intelligenz und Open-Source Intelligence (OSINT) in einzelnen Lektionen abgedeckt.

Das MAZ in Luzern behandelt Deepfakes exemplarisch und im grösseren Zusammenhang von KI-Entwicklungen und der Thematik Desinformation. In der «Diplomausbildung Journalismus» (der Grundausbildung am MAZ) werden in weiterführenden Recherchekursen Täuschungsrisiken und Anforderungen an die Verifikation bzw. Bildforensik diskutiert. Die Thematik wurde auch spezifischer im Rahmen vergangener MAZ-Digitaltage oder Recherchetage behandelt. Im MAZ-Vertiefungskurs «Journalismus» wurden Deepfake-Beispiele in Zusammenhang mit Desinformationsbekämpfung diskutiert. Einige Weiterbildungskurse zu diesem Thema wurden aber mangels Nachfrage nicht durchgeführt. Im CAS «Innovation im Journalismus» schliesslich sind Deepfakes ein (nicht explizit ausgewiesenes) Teilthema des Moduls zu KI. Das MAZ bietet auch eine Linksammlung als Überblick über wichtige Plattformen, Initiativen und sonstige Bemühungen rund um die Bekämpfung von Desinformation an (MAZ 2023).

Die Académie du journalisme et des médias (AJM) der Universität Neuenburg hat uns schriftlich informiert, dass das Thema in mehreren Lehrveranstaltungen integriert wird. Im Kurs «Publizieren, Veröffentlichen und digitale Verwertung» lernen die Studierenden Methoden der Recherche und Verifizierung von Onlineinformationen und das Erkennen von «Fake»-Inhalten wie z.B. Deepfakes. Der Theoriekurs «Information und digitale Medien» befasst sich mit den Herausforderungen von Information und Journalismus im digitalen Zeitalter. Dabei

sind vier Stunden dem Thema «Fake News» gewidmet, mit einem besonderen Schwerpunkt auf Deepfakes. Der Kurs «Visuelle und digitale Kultur» behandelt in einer Sitzung das Thema Deepfakes aus einem etwas technischeren Blickwinkel und in Zusammenhang mit den Fortschritten der künstlichen Intelligenz bei der Generierung von Inhalten. Das Thema wird zudem von den Studierenden des AJM auch als journalistisches Thema für die Website des AJM behandelt (z.B. JAM 2019b oder JAM 2019a). Weiter gibt es Kooperationen zwischen AJM und CFJM. So bietet das CFJM vier berufsbezogene Workshops für Studierende des Masterstudiengangs Journalismus des AJM an.

Am Institut für Angewandte Medienforschung (IAM) der ZHAW fliessen Deepfakes in verschiedenen Bereichen in die Ausbildung mit ein, es wird aber nicht explizit als einzelnes Thema behandelt. Deepfakes werden im Zusammenhang mit Täuschungen, Medienethik und «Kritischem Denken» diskutiert. Thematisiert wird auch, wie mit Bildern Geschichten erzählt werden können, wobei auch auf Grenzen bei der Verwendung von Deepfakes eingegangen wird. Geplant ist ein Kurs «Media Literacy», wo ebenfalls das Thema Deepfakes mit einfliessen soll. Weiter existiert eine eintägige Praxiswerkstatt zum Thema «Bild und Bewegtbild in Social Media», wo unter anderem Deepfakes diskutiert werden. Beim Weiterbildungsangebot «Innovation im Journalismus» gibt es u.a. das Lernziel «Sie sind im Bilde über aktuelle Entwicklungen bei der Automatisation, der Algorithmierung und bei der künstlichen Intelligenz im Journalismus», wobei Deepfakes eine Rolle spielen können.

Hinsichtlich der Forschungsfrage kann zusammengefasst gesagt werden, dass die Herausforderungen, welche Deepfakes für den Journalismus mit sich bringen, in der Ausbildung von Medienschaffenden aktuell vor allem in Zusammenhang mit Desinformation thematisiert werden. Spezifisch zu Deepfakes gibt es nur wenige Kurse. Deepfakes werden eher als Beispiele für die verschiedenen Aspekte von Desinformation und des journalistischen Arbeitens thematisiert.

### 5.3.2. Umgang von Medienorganisationen mit Deepfakes

Die Auswertung der Interviews mit Vertreterinnen und Vertretern von Medienorganisationen erfolgt entlang der im Leitfaden enthaltenen Themenbereiche: Einschätzung von Chancen und Risiken, Prävention innerhalb der Redaktion resp. Medienorganisation, Erkennung und Verhinderung der Weiterverbreitung, Massnahmen nach der Veröffentlichung und Verbreitung durch Medien, Journalistinnen und Journalisten als Betroffene, Sensibilisierung der Nutzenden und Förderung von Medienkompetenz.

#### 5.3.2.1. Einschätzung von Chancen und Risiken

*Unterschiedliche Ansichten, ob Deepfakes das Problem mit Desinformation verstärken:* Deepfakes werden von den Interviewten als Spezialform von Desinformation betrachtet. Nicht speziell Deepfakes, sondern Desinformation allgemein (von den Interviewten oftmals als «Fake News» bezeichnet) und die Folgen für das Vertrauen in die Medien werden als die grössere Herausforderung bezeichnet. Dieser Befund deckt sich mit der Einschätzung der Bevölkerung, wonach Deepfakes als hohes Risiko für den Journalismus eingeschätzt werden: Die befragten Personen befürchten, dass Deepfakes das Vertrauen in Schweizer Medien untergraben und zur Erzeugung von Desinformation verwendet werden könnten (siehe Kapitel 3.3.2). Erwähnt werden insbesondere auch sog. «Cheapfakes» oder «Shallow Fakes»:

> *«Vielleicht ist das nicht die Antwort, die du hören willst, aber Deepfakes sind nicht so weit, dass wir uns jetzt grosse Sorgen machen müssten. Das Problem sind die ‹Shallow Fakes›. Deepfakes benötigen viel Mühe, viel Zeit, viel Engineering, viel Technologie. Aber warum solltest du das tun, wenn du das Gleiche mit Shallow Fakes erreichen kannst, also mit Desinformationskampagnen, die sehr primitiv, aber genauso effektiv sind? Aber Deepfakes werden als Ausrede benutzt, um die Wahrheit zu vertuschen. Trump hat auch behauptet, das Video von ihm, als er im Bus sass und über diese Frau geredet hat [‹Grab them by…›], sei ein Deepfake. Er sagte ‹das ist Fake News, Fake News, Fake News›. Meine Sorge ist weniger, ob es ein Deepfake gibt, in dem Putin sagt: ‹Machen wir einen Bombenangriff auf die USA› oder was auch immer. Mich beschäftigen die anderen Desinformationskampagnen mehr, die auf ein Misstrauen in die Medien setzen.»*[146]

Nicht Deepfakes, «sondern das dauernde Wiederholen, dass Medien fake sind»,[147] und das Verbreiten von Shallow Fakes, die mit einfachsten Werkzeugen hergestellt wurden, seien derzeit für Medienschaffende die Herausforderung. Auch ohne Deepfakes würden viele Menschen den Medien nicht vertrauen: «Wir beobachten jetzt schon oft, dass vielen Artikeln, die wir schreiben, die völlig korrekt sind und für die viel recherchiert wurde, einfach nicht geglaubt wird.

---

[146] Interview TFX110123.

[147] Interview SDE271022.

Da denken viele, dass Journalisten lügen.»[148] Entsprechend solle das Risiko von Deepfakes nicht überschätzt werden.[149] Mehrere Interviewpartner weisen darauf hin, dass es aufwendig sei, gute Deepfakes zu erstellen, und es schwierig vorstellbar sei, «dass dann jeder Laie das perfekte Deepfake produzieren kann und die Journalistinnen und das Medienhaus das Know-how nicht haben, um das irgendwie zu entdecken».[150] Dies deckt sich auch mit im Rahmen der vorliegenden Studie gemachten Erfahrungen zur Produktion hochwertiger Deepfake-Videos.[151]

Inwiefern sich diese Einschätzung seit der Durchführung der Interviews angesichts der öffentlichen Diskussionen um generative KI verändert hat, muss an dieser Stelle offenbleiben. Mehrere Interviewpartner befürchteten, dass sich das Problem von Desinformation mit Deepfakes noch weiter verschärfen könnte, und warnten davor, die technologische Entwicklung zu unterschätzen. «Wenn Deepfake einfacher zu machen sein werden, werden sie auch verbreiteter vorkommen. Das wird das Problem [mit Desinformation] sicher noch vergrössern.»[152] Deshalb müssten Medienorganisationen ihre Mitarbeitenden für das Thema sensibilisieren und technologisch aufrüsten. Die grösste Gefahr wird in jener Phase gesehen, in der eine Technologie so neu ist, dass noch wenig Wissen dazu vorhanden ist.

> *«Immer, wenn eine neue Technologie kommt, wird damit experimentiert und ausprobiert, wie man sie benutzen kann, um die Leute in die Irre zu führen. Das ist eigentlich nichts Neues. Das Risiko liegt glaube ich darin, dass man sich nicht bewusst ist, was alles möglich ist, wenn eine Technologie neu ist.»*[153]

*Ambivalente Beurteilung, ob Deepfakes in der Schweiz zum Problem werden könnten:* Zu Kontakt mit Deepfakes im journalistischen Alltag kam es in den Redaktionen bisher selten. «Des deepfakes qui nous toucheraient directement, qui reprendraient par exemple des images d'une personnalité politique suisse,

---

148 Interview CDW161122.

149 Interview PAG171122.

150 Interview CDW161122.

151 Vgl. https://zenodo.org/records/10847968.

152 Interview 3XC231122.

153 Interview SDE271022.

nous n'y avons pas été confrontés, à ma connaissance.»[154] Mehrere Interviewte führten das auch auf die politische Kultur in der Schweiz und auf die, im Gegensatz zu grossen ausländischen Medien, eher geringe Reichweite und Popularität von Schweizer Medien und deren Journalistinnen und Journalisten zurück. Allerdings schliesst eine interviewte Person nicht aus, dass Deepfakes auch im Schweizer Kontext vorkommen könnten. Die Politik und die Medien seien zunehmend polarisiert und es wäre deshalb «für mich nicht aus der Luft gegriffen, dass irgendwann einmal ein Deepfake auftaucht, in dem etwas gesagt wird, das für die betroffene Person schädlich ist. Gerade bei Menschen oder Journalisten, die sich mit einer eigenen Meinung zu gewissen Themen exponieren, sehe ich eine Riesengefahr.»[155] Mehrere Deepfakes, die seit der Durchführung der Interviews in der Schweizer Politik kursierten, scheinen zu bestätigen, dass Deepfakes auch in der Schweiz zu einem politischen Werkzeug werden (siehe auch Kapitel 6.1.2).

Konfrontiert mit Deepfakes wurden Schweizer Medien bisher in der Auslandsberichterstattung, beispielsweise in Zusammenhang mit dem Krieg Russlands gegen die Ukraine: «Was es gegeben hat, waren Vorfälle im Zusammenhang mit dem Krieg in der Ukraine. Es gab ein Deepfake mit Volodimir Selenskyj, wo es hiess, dass die Ukraine kapituliere, und die Fake-Calls von Klitschko. Also sozusagen indirekt haben wir sicher ein Auge drauf. Gerade im Krieg.»[156] In Fällen, wo Deepfakes in Redaktionen aufgetaucht sind, wurden diese relativ einfach erkannt, wie eine Person erklärt: «Also wir hatten schon effektiv welche, wussten das dann aber oder haben die Bilder von Anfang an als potenzielle Deepfakes behandelt.»[157] Zudem kämen Medien in der Schweiz auch weniger mit Deepfakes in Kontakt, da hierzulande keine grossen internationalen Nachrichtenmedien angesiedelt sind:

> *«Wenn Russland einen Deepfake mit Klitschko publizieren würde, der irgendetwas gegen Putin sagt, dann sind wir sehr wahrscheinlich nicht das erste Medium, das über das Thema berichten würde. Die ersten Medien, die das verifizieren würden, wären die grossen Player, also die Avantgarde wie das CNN-Verifikationsteam, das vierzig Leute umfasst und Bildmaterial mit technologischem Super-*

[154] Interview JOP071222.

[155] Interview HJK091122.

[156] Interview HJK091122.

[157] Interview QWE161122.

*Know-how verifizieren kann. Und erst wenn es die gemacht haben, dann greifen wir das Thema auf.»*[158]

Eine interviewte Person sieht die Risiken von Deepfakes auch eher ausserhalb der Industrienationen in Ländern, in denen Medienschaffende «keinen Zugang zu Technologien [für die Verifikation] haben, vielleicht nicht so gut ausgebildet sind, wo es keine Gegen-Medien gibt, die sagen können: ‹Hey, das stimmt nicht›. Da sehe ich eine grössere Gefahr für Deepfakes.»[159]

Doch es gibt auch die gegenteilige Meinung, dass das Risiko von Deepfakes im Schweizer Journalismus unterschätzt werde. Bisher sei man verschont geblieben, das Bildungsniveau sei hoch, und bisher habe Desinformation in der Schweizer Demokratie keine grosse Rolle gespielt.

*«Je pense que vu que nous on a la chance d'être préservés par rapport à ça, peut-être qu'on a tendance à avoir une vision qui est moins assidue par rapport à ces problèmes dès qu'on traite d'actualité internationale, et justement je pense que c'est important de sensibiliser à ce truc dans les rédactions aussi, simplement pour dire ‹OK, en Suisse, ça se passe plutôt bien›. Mais ça ne fait pas de mal de le rappeler de temps en temps.»*[160]

*Gefahr der Weiterverbreitung nicht erkannter Deepfakes:* In den Interviews wurde auch das Risiko thematisiert, dass Medien Deepfakes unter Umständen nicht erkennen und weiterverbreiten.

*«Je pense que les deepfakes devraient aussi nous conforter dans l'idée que le travail journalistique, ça doit être rigoureux. Ça veut dire multiplicité des sources, ça veut dire vérification multiple accompagnée. Que justement on ait cette petite piqûre de rappel qui nous dit ‹attention› pour des choses qu'on fait peut-être moins bien parce qu'elles sont devenues des automatismes, car voilà quand on fait le même métier depuis des années, il y a des choses qui nous paraissent évidentes. C'est une bonne piqûre de rappel pour dire ‹voilà, on n'est pas à l'abri›, d'autant plus que les technologies*

158 Interview CDW161122.

159 Interview TFX110123.

160 Interview HFK021122.

*sont de plus en plus précises, on n'est pas à l'abri de se tromper quand même, même si on fait bien notre travail.»*[161]

*«Je ne pense pas que les technologies pour modifier en direct soient là, mais c'est sûr que si elles venaient à être développées, ce serait un très grand problème puisque nous faisons beaucoup de vérifications sur la base d'appels téléphoniques.»*[162]

Fehleinschätzungen könnten dabei auch gravierende Folgen haben, speziell wenn es um globale Ereignisse geht.

*«Als zum Beispiel [im Krieg in der Ukraine] eine Rakete in Polen einschlug und man nichts Genaues wusste. Sachen, wo natürlich auch die journalistischen Medien mit der Verbreitung einer Falschinformation, die nicht schnell genug verifiziert wurde, einen Krieg auslösen können. Deshalb glaube ich, dass die grösste Gefahr ist, dass der Journalismus auf etwas reinfällt oder dass man irgendwie in diesem Machtspiel von grossen Playern manipuliert wird. Das kann dann fatale Auswirkungen haben.»*[163]

*Gefahr der Manipulation von Medieninhalten:* Ein weiteres Risiko wird von den Interviewten darin gesehen, dass eigene Medieninhalte für ein Deepfake verändert und missbraucht werden. Dieses Risiko reiht sich in den bereits vorkommenden Missbrauch von Medieninhalten für Desinformation ein. Schon heute werden journalistische Texte und Fotos manipuliert oder aus dem Kontext gerissen. Aus Medieninhalten erstellte Deepfakes wären folglich der nächste technische Schritt im Missbrauch von Journalismus für Desinformation: «Das wäre dann noch eine neue Qualität [der Desinformation]. Es ist mir nicht bekannt, dass so etwas schon passiert ist, aber ja, das ist keine schöne Vorstellung. Ich glaube, es gibt nichts, womit wir nicht rechnen sollten. Gerade auch wenn es darum geht, uns bewusst zu diskreditieren.»[164] Deepfakes wird ein grösseres Manipulationspotenzial zugesprochen als anderen Formen des Missbrauchs von Medieninhalten.

---

[161] Interview HFK021122.

[162] Interview PAG171122.

[163] Interview CDW161122.

[164] Interview QWE161122.

*«C'est un risque qui s'ajouterait à des risques qui existent déjà aujourd'hui, et quelque part le risque le plus fort c'est qu'une vidéo où on voit quelqu'un qui parle ait l'air d'être plus vraie qu'une photo ou un détournement habituel. Et il y a ce côté où on se dit que la vidéo, parce qu'on la voit, c'est que c'est vrai.»*[165]

Allerdings relativieren mehrere Interviewpartner das Risiko, das nun durch Deepfakes dazukommt.

*«Je pense que le deepfake reste quand même une technologie qui n'est pas complètement mature d'une part et d'autre part qui est pas à la portée de tout le monde, donc voilà on pourrait imaginer que ce serait beaucoup plus simple de le faire avec des moyens comme simplement changer le titre d'un article, d'un journal ou etc.»*[166]

Ähnlich sehen das auch Vertreter des TV-Journalismus. «Ich kann mir nicht vorstellen, dass sich Leute die Zeit nehmen, wirklich gute Deepfakes mit unserem Videomaterial herzustellen. Wenn, dann vielleicht mehr oberflächlich.»[167]

*Negative Auswirkungen von Deepfakes auf Medienvertrauen befürchtet:* Sowohl die versehentliche Verbreitung von Deepfakes durch die Medien selbst wie auch die Diskreditierung von Medienorganisationen durch Dritte bergen das Risiko des Vertrauensverlusts gegenüber den Medien. Man müsse auf die journalistische Arbeit achten, denn die Leserinnen und Leser würden erwarten, dass die Informationen gut recherchiert und korrekt sind und dass sie entsprechend analysiert und verifiziert werden.[168] Dabei seien Glaubwürdigkeit und Vertrauen im Journalismus zentral. «Das Einzige, was wir haben, ist unsere Glaubwürdigkeit.»[169] Anders als bei sonstigen Formen der Desinformation sehen einige Interviewpartner speziell bei Deepfakes die Gefahr, dass Nutzerinnen und Nutzer nicht mehr zwischen wahr und falsch unterscheiden können. «Viele Kinder, aber auch Jugendliche und sicher Erwachsene, die ein Deepfake auf einer Plattform sehen, glauben, dass es richtig ist, und hinterfragen das Video nicht kritisch.

165 Interview JOP071222.

166 Interview PAG171122.

167 Interview I2 3XC231122.

168 Interview HFK021122.

169 Interview TFX110123.

Man kann Videos oder Bilder so schnell verbreiten, auf der ganzen Welt. Und das ist natürlich auch eine Gefahr.»[170] Dies evoziert bei einigen dystopische Vorstellungen einer Zukunft, in der «man am Schluss nicht mehr weiss, was man glauben kann und was nicht. Das wäre eine Katastrophe und das muss man vermeiden».[171] Trotzdem weisen viele der Interviewten wieder darauf hin, dass die Technologie dafür noch nicht ausreiche.

> *«Dans un monde où on aurait des deepfakes ultra bien faits, où tout le monde pourrait en fabriquer on serait inondés, sans doute que ça serait compliqué, il faudrait que nous les médias on s'interroge sur notre manière de faire les choses. Par contre dans le monde dans lequel on est aujourd'hui, où cette technologie n'est pas encore accessible, ou trouve malgré tout il y a des citoyens qui remettent en question toute vérité et certitude parce que finalement la complexité du monde fait aussi que ça ne permet plus de différencier ce qui est vrai, ce qui est faux.»*[172]

*Skeptische Einschätzung von Chancen für den Journalismus:* Spezifisch auf die Chancen von Deepfake-Technologien für den Journalismus angesprochen, überwog bei allen Interviewten zuerst einmal Skepsis. Das Ergebnis reiht sich ein in den auch in der Bevölkerung (siehe Kapitel 3) und der Politik (siehe Kapitel 6.2) vorherrschenden Fokus auf Risiken von Deepfakes. Die skeptische Haltung der befragten Journalistinnen und Journalisten gegenüber Deepfakes dürfte damit zusammenhängen, dass im journalistischen Kontext Deepfakes als technisch versierte Form von Desinformation betrachtet werden und es im Idealfall Aufgabe des Journalismus ist, gemäss professioneller Normen verifizierte Informationen bereitzustellen. Entsprechend werden Deepfakes als Versuch gesehen, die erfolgreiche Erfüllung dieser Aufgabe zu unterlaufen. «Das Risiko für uns Medien halte ich vielleicht aufgrund meiner beruflichen Brille für grösser als das Potenzial.»[173] Viele Interviewpartner hatten deshalb Mühe, positive Aspekte für den Journalismus zu sehen: «Finalement, l'objectif du journalisme, c'est la

[170] Interview 3XC231122.

[171] Interview CDW161122.

[172] Interview JOP071222.

[173] Interview QWE161122.

recherche de la vérité, hors cette technologie qui permet de modifier la réalité et de créer une vérité alternative, ça va à l'encontre de cet objectif.»[174]

Doch auch jenseits der Verantwortung des Journalismus, Deepfakes zu erkennen, zeigten sich die Interviewten kritisch gegenüber dem Potenzial von Deepfake-Technologien für den Journalismus. Zu gross seien die Gefahren für die eigene Glaubwürdigkeit. Einige sehen die Produktion von Deepfakes durch die Medien selbst als Widerspruch zu ihrem Beruf und zu journalistischen Leitwerten.

> *«Donc j'espère qu'on n'en n'arrive pas là, moi je crois encore à la bonne foi et puis au travail rigoureux de nous journalistes, sinon je ne ferais pas ce métier. Mais alors après c'est comme toutes les technologies, ça peut être à la fois des outils fabuleux qui nous permettent de faire des choses, etc. Mais quand c'est mal utilisé, ou dans les mauvais moments, ça devient très vite nocif.»*[175]

Auch im Videojournalismus gelte, dass Bilder so gezeigt werden müssten, wie sie sind: «Nicht verändern, beziehungsweise klar einordnen, wenn man nicht weiss, ob Bilder real sind. Bei Deepfakes in Newsbeiträgen bin ich skeptisch, weil dort Faktizität über allem steht.»[176] Weiter müsse Journalismus Desinformation entlarven und nicht selbst produzieren.[177] Gerade in Anbetracht der bereits heute angespannten Situation für den Journalismus aufgrund von Desinformation sei hier Vorsicht angezeigt: «Dans le contexte des réseaux sociaux, il est important de se positionner comme défenseur de la vérité, de la recherche des faits, de l'exactitude des faits. Et donc si nous utilisions des deepfakes et fake news nous-même, ce serait un problème, on peut risquer de se brûler les doigts en quelque sorte.»[178]

Viele Aussagen zu den potenziellen Chancen von Deepfake-Technologien für den Journalismus weisen auf ein Spannungsverhältnis zwischen technischer Entwicklung und journalistischen Grundwerten hin. «Moi je suis plutôt partisan du travail avec les technologies, du travail avec les intelligences artificielles et les algorithmes, on doit toujours bien les utiliser, mais le deepfake c'est quel-

---

[174] Interview JOP071222.

[175] Interview HFK021122.

[176] Interview HJK091122.

[177] Interview HJK091122.

[178] Interview JOP071222.

que chose d'un peu différent quand même.»[179] Viele sehen die Gefahr, dass die Skepsis in der Gesellschaft gegenüber Medien dadurch noch gefördert werden könnte. «Ich glaube, man muss vermeiden, dass die Nutzerinnen und Nutzer nicht mehr wissen, was echt ist und was fake.»[180] Ebenso wird auf den grundsätzlichen Stellenwert von Journalismus für die Gesellschaft hingewiesen und betont, dass dazu auch eine kritische Haltung gegenüber Technik und synthetischen Medienprodukten gehören sollte.[181]

*Chancen für den Journalismus:* Trotzdem finden sich in den Interviews auch Ideen, wie die Technologie hinter Deepfakes unter bestimmten Bedingungen im Journalismus genutzt werden könnte. Wie auch die Teilstudie zu Deepfakes in der Wirtschaft zeigt (siehe Kapitel 7.2), werden die Chancen von Deepfakes vor allem in Bezug auf Unterhaltung oder bildende Inhalte gesehen. Zum Beispiel wird der Nutzen für Erklärvideos diskutiert. «So wie früher Grafiken gemacht wurden, könnte man jetzt Erklärvideos machen. Aber es müsste auch wirklich offensichtlich sein. Der Leser dürfte nicht reingelegt werden.»[182]

Zudem wurde auch ein spezifischer Nutzen von Deepfakes für verdeckte Recherchen erwähnt. Diese würden durch synthetisch erstellte Bilder oder Videos klar vereinfacht.

> *«Was ich mir halt auch überlege, ist, dass Deepfakes bei Recherchen vielleicht sogar eine Chance sein könnten. Also jetzt auf mich persönlich bezogen, weil ich viele verdeckte Recherche mache. Da brauche ich z.B. KI-generierte Fotos, wenn ich mit falschen Profilen auf Facebook arbeite, zum Beispiel im Bereich des Rechtsextremismus. Da habe ich das schon vor zehn, fünfzehn Jahren gemacht. Und als man ein Profil erstellen wollte, brauchte man ein Foto. Das ist natürlich sehr heikel, wenn man da echte Personen nehmen muss. Und heute kann man mit KI-generierten Fotos Personen erzeugen, die es schlichtweg nicht gibt, und das relativ gut.»*[183]

[179] Interview JOP071222.

[180] Interview CDW161122.

[181] Interview TFX110123.

[182] Interview CDW161122.

[183] Interview CDW161122.

Mit den hinter Deepfakes stehenden Technologien wäre es weiter möglich, die Inhalte von Nachrichten zu personalisieren und Avatare von Moderatorinnen und Moderatoren zu erstellen. In der Schweiz gibt es Beispiele von komplett synthetischen Moderatoren (Widmer 2023) oder Sendungen, deren Inhalte mithilfe von künstlicher Intelligenz produziert und von synthetischen Moderatorinnen geführt wurden (z.B. RTS Info 2023). Für solche Anwendungen sehen die Interviewten schon eher Potenzial im Journalismus.

> *«Wir haben mit der ETH Zürich zwei Avatare entwickelt, einen Moderator und eine Moderatorin, mit denen Deepfake-Videos gemacht werden können. Die Technologie ist da, solche Videos gibt es im Netz und wir müssen uns mit diesem Thema befassen. Wir haben jetzt aktuell keine konkreten Use-Cases, sonst hätten wir auch irgendwo ein Vertrauensproblem, denn die Leute erwarten ja auch, dass ein richtiger Mensch am Mikrofon oder vor der Kamera steht. Und ja, wir müssen uns mit dieser Technologie befassen, denn sie ist da und wir müssen uns auch überlegen, wo es Chancen für uns gibt, aber auch wo Risiken für uns bestehen.»*[184]

Personalisierung sei «grundsätzlich immer ein Plus».[185] Aber auch hier brauche es Transparenz.[186] Andere sehen Limitationen bei der noch jungen Technik und der vermuteten geringen Akzeptanz beim Publikum.

Potenzial wird auch darin gesehen, dass Avatare «als Hilfestellung auf der Website zur Erklärung eingesetzt werden könnten. Es gibt im Bereich der Accessibility-Services Möglichkeiten, aber so wie ich das momentan einschätze, ist es einfach noch zu früh. Im Moment sind die Risiken noch grösser als die Chancen.»[187]

Auch die Möglichkeit, Moderatorinnen und Moderatoren oder Aussagen von Personen in Nachrichtenbeiträgen automatisch in verschiedenen Sprachen zu übersetzen, stiess bei vielen der Interviewten auf positive Resonanz. «Effectivement, ça peut être un truc qui est intéressant pour toucher une plus large audience.»[188] Allerdings gibt eine Person zu bedenken, dass bei der heutigen

---

[184] Interview 3XC231122.

[185] Interview HJK091122.

[186] Interview QWE161122.

[187] Interview 3XC231122.

[188] Interview HFK021122.

Lösung mit Dolmetschen immerhin die Originalsprache auch noch hörbar ist, was bei einer Automatisierung entfallen würde: «Quand on n'entend pas tout ce qu'ils disent et quand ils sont doublés en direct par-dessus, les gens se disent ‹est-ce que ce ne serait pas truqué?› Donc je ne serais pas convaincu.»[189] Auch beim öffentlichen Rundfunk wird diese Meinung geteilt: «Die Leute haben Vertrauen in uns und mit dem sollte nicht gespielt werden. Ein privates Medium hat da eventuell mehr Möglichkeiten, einen Avatar einzusetzen oder Stimmen auszutauschen als wir.»[190] Aber auch bei den privaten Medien gibt es kritische Stimmen:

> *«Ich finde es sogar bei Übersetzungen heikel, weil es ja quasi auch ein Fake ist. Beispielsweise wenn dann der Moderator zur türkischen Community plötzlich auf Türkisch spricht, aber es eigentlich ein Fake ist. Da muss man einfach sehr aufpassen, dass die Konsumentinnen und Konsumenten nicht irgendwann nicht mehr nachkommen. Aber [automatisierte] Sprachen, da sehe ich Deepfakes jetzt noch am ehesten.»*[191]

Auch weitere Interviewpartner sind skeptisch: «Das ist ein alter Traum von der Automatisierung der journalistischen Produktion. Der wird auch mit Deepfake-Technologien nicht in Erfüllung gehen.»[192]

Nachgedacht wurde auch über die Nutzung synthetischer Medienprodukte wie Deepfakes zum Nachstellen bestimmter Ereignisse in Dokumentarfilmen:

> *«Le problème que je verrais c'est que pour les journalistes, si on utilise des deepfakes qui reflètent la vérité, eh bien on est hors du cadre qui est le cadre des journalistes, parce que même si vous avez les preuves pour vous que ça s'est passé comme ça, si vous inventez des images ou créez des images de ce qui s'est passé, vous créez quelque chose de faux. Alors j'ai beaucoup de peine à voir des aspects positifs.»*[193]

---

189 Interview JOP071222.

190 Interview 3XC231122.

191 Interview CDW161122.

192 Interview SDE271022.

193 Interview JOP071222.

Die im Forschungsstand thematisierte Möglichkeit, verstorbene Personen sprechend in Videos darstellen zu können, lehnen die meisten Interviewten vehement ab. «Mais en fait c'est justement là qu'on sort du journalisme.»[194]

Letztlich wollen viele Interviewte trotz kritischer Haltung neue Trends und Chancen nicht verpassen. «Also sehr wahrscheinlich müssen wir mit diesem technologischen Fortschritt mitgehen. Man sollte sich nicht grundsätzlich verschliessen, sondern auch die Chancen ins Auge fassen.»[195] Ein Nutzen synthetischer Medien wird aber vor allem ausserhalb des Journalismus gesehen, insbesondere im Bereich der Unterhaltung und der Satire.

> *«Wenn, dann im satirischen Bereich. Also wenn man vor der Herausforderung steht, Themen, die objektiv langweilig sind wie Politik, für die Leute subjektiv relevant zu machen. Hier könntest du natürlich eine alternative Storytelling-Form einsetzen, die Leute an den Content heranführen und den Content zu den Leuten führen auf diesen Kanälen und auf diese Geräte, wo sie sonst keine News oder objektiv relevanten Inhalte bekommen würden.»*[196]

*Erfolgreiche Erkennung von Deepfakes als Qualitätsmerkmal:* Eine grosse Chance sehen viele Interviewte darin, dass Medien ihren erfolgreichen Umgang mit Deepfakes als Qualitätsmerkmal herausstreichen können. Wenn Medien ihre Aufgabe der verantwortungsvollen Verifikation von Informationen erfüllen und dem Publikum aufzeigen können, was Deepfakes sind und wie man diese erkennen kann, dann könne die Rolle der Medien in der Gesellschaft gestärkt werden.[197] Dabei geht es nach Ansicht vieler Interviewter nicht nur um das Erkennen von Deepfakes, sondern auch um das Thematisieren und Sensibilisieren des Publikums, das eben letztlich auch als Chance für den Journalismus gesehen wird.

> *«Vor ein paar Wochen, als es zu Raketenangriffen auf Kiew kam, ging ein Video mit einer jungen Frau viral, das zeigte, wie hinter ihr eine Rakete direkt einschlägt. Dort konnten wir anhand verschiedener Visual Investigation Tools verifizieren, dass es dort so stattgefunden hat, mit 99.9-prozentiger Wahrscheinlichkeit. Ein*

---

194 Interview JOP071222.

195 Interview CDW161122.

196 Interview SDE271022.

197 Interview CDW161122.

*kleiner Rest an Gefahr bleibt natürlich immer. Was wir dann auch transparent machen.»*[198]

Insgesamt werde die journalistische Arbeit dadurch wertvoller, dass es eine grosse Menge an Desinformation gibt – zumindest wenn Medien glaubwürdig aufzeigen können, dass ihre Inhalte vertrauenswürdig sind. «Le problème est qu'ensuite les croyances des gens s'organisent souvent en fonction de pré-croyances. Même si on veut faire en quelque sorte un journalisme impartial, on sera toujours accusés, si ça ne va pas dans le sens de ce que les gens voulaient déjà penser, ils diront que c'est idéologique, ou c'est mensonger. C'est une très grande difficulté de notre travail aujourd'hui.»[199]

#### 5.3.2.2. Prävention innerhalb der Redaktion resp. Medienorganisation

*Bewusstsein für Deepfakes stärken:* Für eine gelingende Prävention der unabsichtlichen Verbreitung von Deepfakes in Redaktionen und Medienorganisation ist es wichtig, dass das Bewusstsein für Risiken von Deepfakes vorhanden ist. «Der entscheidende Punkt ist: Bist du dir bewusst, dass es das gibt? Wenn du dir dessen nicht bewusst bist, fällst du bei einem Deepfake rein.»[200] Dieses Bewusstsein sollte aber bereits in allen Redaktionen vorhanden sein. «Wir sind uns alle bewusst, dass es das gibt.»[201] Dabei geht es nicht unbedingt darum, genau zu wissen, wie Deepfakes entlarvt werden können, sondern eher die Skepsis zu fördern und sich der journalistischen Standards bei der Verifikation von Information bewusst zu sein. Ein grundsätzliches Bewusstsein für Deepfakes und eine skeptische Haltung bei der Verifikation von Information sollte eigentlich bei allen Journalistinnen und Journalisten vorhanden sein. In schwierigen Fällen müssen dann aber spezialisierte Abteilungen und Personen zugezogen werden. «Die Knacknüsse werden an die Taskforce weitergegeben, wo dann halt wirklich die Freaks sitzen.»[202] Aus den Aussagen der Interviewten geht hervor, dass in den Redaktionen aufgrund der zunehmenden Konfrontation mit Desinformation be-

198 Interview HJK091122.

199 Interview PAG171122.

200 Interview SDE271022.

201 Interview QWE161122.

202 Interview QWE161122.

reits ein starker Fokus auf Verifikationsmechanismen liegt und die Journalisten diese bereits gut internalisiert haben.[203]

*Sensibilisierung durch spezifische Ereignisse:* Oft wurde in den Gesprächen erwähnt, dass die Sensibilisierung für Deepfakes im journalistischen Alltag geschieht. «C'est devenu un grand sujet, auquel tout le monde est sensibilisé. Ça s'est fait de façon beaucoup plus naturelle.»[204] Wenn Journalistinnen und Journalisten einmal auf ein Deepfake gestossen oder darauf hereingefallen sind, sind sie vorsichtiger. «Das geschieht dir einmal und danach bist du relativ gut sensibilisiert.»[205] Andere erinnern daran, dass man in der journalistischen Praxis die Evolution mitverfolgen konnte und daher durch einfacher zu entlarvende Deepfakes auch für schwierigere Fälle sensibilisiert wurde. «Ich bin dankbar dafür, dass es ein paar Deepfakes gegeben hat, die entweder aufgedeckt wurden oder bei denen kommuniziert wurde ‹Das ist jetzt eins›.»[206] Viele Redaktionen hätten Deepfakes auch in der Berichterstattung zum Thema gemacht und dadurch gleich die beteiligten Journalistinnen und Journalisten für das Thema sensibilisiert.

> *«La meilleure manière d'en parler, c'est encore de le documenter. Il y a eu quelques reportages. Notamment, on a aussi fait un reportage pour créer un deepfake, pour voir de A à Z quelle technologie, combien ça coûte, comment on l'obtient, qu'est-ce qu'on peut avoir à la fin, dans quel but? C'était une manière de faire un peu un état des lieux, d'où on en était avec ça.»*[207]

*Schulung und Weiterbildung in Redaktionen oder Medienorganisationen:* Spezifische Schulungen und Weiterbildungen für den Umgang mit Deepfakes gibt es in den meisten Redaktionen nicht. «Les fake news sont effectivement quelque chose qui occupent beaucoup les rédactions, mais pas les deepfakes. En tout cas pas pour le moment.»[208] Schulungen seien nicht Deepfake-spezifisch, son-

[203] Interview SDE271022.

[204] Interview PAG171122.

[205] Interview SDE271022.

[206] Interview QWE161122.

[207] Interview JOP071222.

[208] Interview PAG171122.

dern es gehe um Manipulationen generell.[209] Wenn also auch keine Schulungen speziell für den Umgang mit Deepfakes existieren: Verifikation und Desinformation sind intern wichtige Themen für die Weiterbildung und ein zentrales Element im journalistischen Alltag geworden. «Das ist natürlich Dauerthema bei uns. Da gibt es Kurse und Schulungen für Verifikationen, da sensibilisiert man die Leute. Das ist ja auch ein Stück weit das journalistische Einmaleins.»[210]

Aus den Einschätzungen der Interviewten geht auch hervor, dass Open-Source Data und Open-Source Intelligence (OSINT) wichtige Instrumente sind, die bei der Erkennung von Deepfakes helfen können. Unter OSINT wird das Sammeln und Analysieren von Informationen aus offenen, online verfügbaren Daten und Quellen verstanden, um daraus nützliche Erkenntnisse zu gewinnen und Rückschlüsse auf die Echtheit von Informationen wie Bildern und Videos zu machen (European Commission 2022). In den Interviews wurde dazu unter anderem auf das investigative Kollektiv Bellingcat (www.bellingcat.com) verwiesen, welches mittels OSINT recherchiert. «Wir haben auf eigene Initiative hin den Kurs ‹Video verification› absolviert. Da ging es nicht nur um Deepfakes, sondern auch sehr viel um ‹Open-Source Data›.»[211] In vielen Redaktionen werden die Journalistinnen und Journalisten zumindest mit Grundwissen geschult. «Alle unsere Redakteure werden drei Tage lang im Bereich ‹Open-Source Intelligence› ausgebildet. Wie man recherchiert, wie man Deepfakes erkennt.»[212] Dadurch sollen sie «das Grundprinzip von einem Faktencheck mit OSINT»[213] lernen. Damit komme man im Alltag schon sehr weit. Die «harten Fälle»[214] übernehmen dann spezifische Fachleute oder Abteilungen, die häufiger mit manipuliertem Video- und Bildmaterial konfrontiert sind. «Wenn bei mir beim Videodesk neue Mitarbeiterinnen und Mitarbeiter anfangen, dann habe ich Schulungen mit ihnen, wo ich ihnen auch sage, dass das ein Teil ihrer Arbeit ist.»[215]

[209] Interview SDE271022.

[210] Interview CDW161122.

[211] Interview HJK091122.

[212] Interview TFX110123.

[213] Interview 3XC231122.

[214] Interview QWE161122.

[215] Interview HJK091122.

#### 5.3.2.3. Erkennung und Verhinderung der Weiterverbreitung

*Journalistische Standards für die Verifikation:* Die meisten Interviewten waren der Meinung, dass Deepfakes im Moment noch recht einfach erkennbar seien. «Deepfakes sind meiner Meinung nach noch nicht so weit, dass wir als erfahrene Journalisten das nicht erkennen können.»[216] Entsprechend wurde wiederholt darauf hingewiesen, dass bei der Erkennung von Deepfakes grundsätzlich die herkömmlichen journalistischen Standards und Verifikationsstrategien entscheidend seien.[217] «Also die ganzen journalistischen Grundprinzipien, wie das Zwei-Quellen-Prinzip oder dass man User-Generated Content überprüft, bevor er benutzt wird. Das wird alles praktiziert bei uns.»[218] Hierfür würden Journalistinnen und Journalisten ausgebildet. «Il y a ce processus-là. Et du coup, de ce processus qu'on apprend tout de suite à l'école, découle automatiquement une forme de pensée et d'analyse qui va mener la réflexion plus loin. Mais en matière de deepfakes et le fait de reconnaître le deepfake, il n'y a aucune mesure en tant telle qui est prise.»[219] Diese Prozesse sollten nach Ansicht vieler Interviewter in der Arbeitspraxis internalisiert sein. «Das Wichtigste ist immer noch der gesunde Menschenverstand. Damit kann man schon 99 Prozent der Fakes aussieben.»[220]

Als weiteren Aspekt bei der Erkennung und Verhinderung der Weiterverbreitung nannten viele Interviewte auch, dass man sich an anderen Medien orientiert. «Wenn Deepfakes auftauchen, dann werden sie sofort von einem anderen Medium als das entlarvt, was sie sind.»[221] Dabei stützen sich viele auch auf bereits verifizierte Quellen, wie z.B. Material von Nachrichtenagenturen. Ein Risiko bestehe vor allem bei unbekannten Quellen aus dem Internet.

> *«Il faut toujours trouver un moyen de vérifier son image, sa source, mais comme on le fait pour d'autres choses. Donc finalement ça rajoute une complexité dans un travail de filtrage que l'on fait déjà au quotidien. Il faut juste s'en rappeler, avoir des bons réflexes.*

---

[216] Interview TFX110123.

[217] Interview SDE271022.

[218] Interview 3XC231122.

[219] Interview HFK021122.

[220] Interview CDW161122.

[221] Interview 3XC231122.

*Mais c'est notre travail journalistique de vérifier les images que l'on partage.»*[222]

Gerade bei ausländischen Quellen kann das schwieriger sein. Dies wiederum unterstreicht den Stellenwert von Auslandskorrespondentinnen und Partnern vor Ort. «Ich glaube, die beste Variante ist immer noch, wenn man People on the ground hat. Also wenn es irgendein offizieller Event ist, ist die Wahrscheinlichkeit relativ gross, dass ein Journalist von uns dort ist. Sonst gilt es schlicht und einfach, die Information zu verifizieren.»[223]

Andere Interviewte betonten allerdings, dass der Faktencheck bei Deepfakes aufgrund deren Komplexität «eine Stufe weiter gehe».[224]

*«Ich nenne unsere Leute, böse gesagt, ‹bullshit detectors›. Bei jedem Video, das bei uns reinkommt, sollen sie zuerst mal davon ausgehen, dass es fake ist. Vor allem, wenn es etwas ist, dass dich besonders aufwühlt oder besonders unglaublich ist. Wo wurde es aufgenommen? Ist es verifiziert? Und zwar nicht nur von zwei, drei unbekannten Twitter-Leuten, sondern von einem Korrespondenten eines grossen Leitmediums. Und auch wenn dieser sagt, dass es so passiert ist, ist das für mich nicht ein Beweis, aber man kann sich dann im Video darauf beziehen: ‹Er sagt, diese Quelle sei vertrauenswürdig, wir selbst können es aber nicht zu 100 Prozent verifizieren›. Also eigentlich machen wir immer transparent, woher ein Video kommt.»*[225]

*Trend weg vom schnellen Journalismus:* Viele Interviewte betonten, dass die Verifikation von Information in den letzten Jahren noch wichtiger geworden sei. Glaubwürdigkeit und Faktenorientierung gehörten gerade im Kontext von Desinformation zu den wichtigsten journalistischen Standards. Allerdings brauchen Verifikationsprozesse Zeit, die im Journalismus knapp ist. Dabei sehen einige Interviewte auch eine Trendwende.

---

[222] Interview JOP071222.

[223] Interview HJK091122.

[224] Interview QWE161122.

[225] Interview HJK091122.

*«Früher ging es mehr um Tempo. Jetzt ist es eigentlich das Gegenteil. Ein zentraler Pfeiler gegen Desinformation und Fakes ist auch eine gewisse Entschleunigung. Besser einmal mehr nachdenken, bevor ein Push herausgeht. Ein wenig Tempo rausnehmen.»*[226]

Das Ziel sei «mehr Qualität, mehr Einordnen, mehr Abheben von ‹alternativen› Medien»,[227] um Verlässlichkeit zu bieten und die eigene Verantwortung wahrzunehmen.

*Spezialisierte Arbeitsgruppen und Abteilungen:* Nach Ansicht vieler Interviewter sind Deepfakes vor allem ein Thema für Abteilungen, die sich bereits jetzt hauptsächlich mit Informationen aus dem Internet beschäftigen. «Die Social-Media-Abteilung und die Video-Abteilung sind halt unglaublich auf das geschult. Also die können das. Die können dir ohne Probleme ein Deepfake machen und die können dir auch sagen, wie du erkennst, ob es ein Deepfake ist.»[228] Auch externe Experten können im Zweifelsfall kontaktiert werden. «Wenn ich jetzt die Situation hätte, wo ich mir nicht sicher wäre, ob es ein Deepfake ist und es wirklich sehr wichtig wäre, dass wir das Video aufgrund einer Geschichte zeigen möchten, dann hätte ich meine Expertinnen und Experten ausserhalb des Unternehmens, die ich fragen könnte, ob sie irgendwelche Tools haben, mit denen sie mir helfen können.»[229]

Allerdings scheint der Bedarf an Expertise zuzunehmen, was sich auf die Belastung von einzelnen Personen auswirkt. «Bis jetzt sind das Journalisten und Dokumentalisten, die ein grosses Wissen mit Faktenchecking besitzen, aber wir machen das quasi alle neben unserem täglichen Geschäft.»[230] Ein Spezialist meinte dann auch, «ich bin permanent ausgebucht mit Anfragen von Kollegen, um eine Situation zu analysieren. Die kommen immer zu mir, wenn irgendwas nicht plausibel klingt. Nicht, dass ich alles weiss, aber wenn ich es nicht weiss, weiss ich, wo ich die Antwort finde.»[231] Deshalb gibt es vor allem bei grösseren Medienorganisationen Initiativen, um spezialisierte Abteilungen aufzubauen.

---

[226] Interview CDW161122.

[227] Interview QWE161122.

[228] Interview SDE271022.

[229] Interview HJK091122.

[230] Interview 3XC231122.

[231] Interview TFX110123.

Diese sollen aber nicht nur Deepfakes, sondern generell Fakten überprüfen oder für neue Technologien zuständig sein. «Wir haben diese Taskforce zur Video- und Fotoverifikation gegründet. Diese beschäftigt sich auch mit Deepfakes. Ein sehr kleines, spezialisiertes Team, was so ein bisschen nerdig unterwegs ist für so diese ganzen Feinheiten.»[232] «Zurzeit bestehen interne Arbeitsgruppen, die das Potenzial und die Gefahren von Künstlicher Intelligenz im Allgemeinen zum Thema haben.»[233] Der Aufbau von auf Deepfakes spezialisierten Arbeitsgruppen oder Abteilungen lohnt sich aber nicht für jede Medienorganisation. Das Budget ist bei kleineren Medien dafür meistens zu knapp. «Man muss sich darüber im Klaren sein, dass es sich um Unternehmen handelt, die in einem ziemlich angespannten wirtschaftlichen Umfeld agieren.»[234]

*Kooperation mit anderen Medienorganisationen:* Kooperationen zwischen Medien scheinen im Bereich der Verifikation und Austausch von Information eine gute Strategie zu sein. Dabei kann bei Material von Partnern auf deren Verifikation vertraut werden.

> *«Il y a des sources où on est 100% sûrs de leur véracité, c'est ce que nous-mêmes on a tourné, nos images. Ensuite il y a des sources où on est très sûrs aussi, on ne peut pas dire à 100%, sans doute parce que ce ne sont pas nos images, mais ce sont des images de nos partenaires : ce sont des images de SRF et RSI par exemple, ou des images d'autres unités d'entreprises. Quand on reçoit des images des membres de l'Union européenne de radiodiffusion l'EBU, on sait que les images que les personnes nous envoient, ce sont des images qui sont comme celles que nous envoyons. Et les images d'agences aussi, il y a un contrat de confiance.»*[235]

Neben formellen Kooperationen zwischen Medienorganisationen finden sich auch informelle Formen der Zusammenarbeit zwischen einzelnen Journalistinnen und Journalisten. «Wir helfen uns gegenseitig, weil man darf nicht vergessen, als Medienhäuser sitzen wir alle im selben Boot und wir wollen alle einen

---

[232] Interview QWE161122.

[233] Interview CDW161122.

[234] Interview PAG171122.

[235] Interview JOP071222.

guten Ruf für den Journalismus.»[236] Zudem wurde erwähnt, dass bei der Verifikation von Informationen auch auf Communities zurückgegriffen werden kann. «Es gibt eine recht starke Community auf Twitter, wo ich nachfragen kann, wenn ich nicht weiterkomme.»[237] Dabei wünschen sich viele Interviewte, dass Kooperationen und gemeinsame Initiativen weiter ausgebaut würden, «weil wir sind ja eigentlich alle mit den gleichen Herausforderungen konfrontiert».[238] Gerade wenn es um technisches Know-how geht, sei eine Kooperation wichtig, da hier häufig wenig Ressourcen vorhanden seien.[239]

*Filtertechniken:* Angesprochen auf den Nutzen spezifischer Software und Filtertechniken für die Erkennung von Deepfakes fielen die Einschätzungen unterschiedlich aus. Einige Interviewte finden Filtertechniken perspektivisch wichtig.[240] Dabei wurde aber auch darauf verwiesen, dass in vielen Redaktionen ein vertieftes Wissen über die Technologie fehlt. Dies auch deshalb, weil Deepfakes bisher selten waren. «Ich musste noch nie einen Deepfake validieren, zum Glück, weil uns eben dieses Wissen fehlen würde. Also wir machen natürlich Fake-Checks, eben auch von Videos und Fotos. Aber technologisch Deepfakes auseinandernehmen, das können wir nicht.»[241] Allerdings wird auch in diesem Zusammenhang auf die beschränkten finanziellen Mittel und die Verhältnismässigkeit hingewiesen. «Also die New York Times wird mit ihrer Videoforensik auf der Höhe sein. Aber für Schweizer Verlage ist das nicht interessant. Da würde man mit Kanonen auf Spatzen schiessen.»[242] Auch weitere Interviewte stimmten zu, dass die bisherigen Mittel für Deepfakes ausreichend seien.

> *«Bis jetzt haben wir kein Problem mit Deepfakes gehabt und von daher sind wir super sensibilisiert und bis jetzt kann man das einfach mit einer Untersuchung erkennen, ob etwas ein Deepfake ist oder nicht. Ich glaube aber, wir werden uns in den nächsten Jahren mehr mit diesem Thema beschäftigen müssen.»*[243]

---

[236] Interview TFX110123.

[237] Interview QWE161122.

[238] Interview QWE161122.

[239] Interview CDW161122.

[240] Interview QWE161122.

[241] Interview 3XC231122.

[242] Interview SDE271022.

[243] Interview TFX110123.

#### 5.3.2.4. Massnahmen nach der Veröffentlichung und Verbreitung durch Medien

Trotz allen Strategien zur Erkennung von Deepfakes und Verifikation von Information kann eine versehentliche Verbreitung nie ganz ausgeschlossen werden. «Es kommen Fehler vor. Auch bei uns arbeiten nur Menschen.»[244] Falls ein Deepfake verbreitet würde, ist für die Interviewten klar, dass man versuchen müsste, schnell zu reagieren, um «möglichst rasch in die Gegenrichtung zu steuern».[245] Dabei komme es darauf an, über welche Kanäle das Deepfake verbreitete wurde.

> *«Alors pour nous la stratégie elle est très claire, c'est que dès le moment où il y a une erreur éditoriale qui a été faite, c'est qu'on va la reconnaitre et on va l'assumer en tant que telle. Aujourd'hui, avec le numérique, ça serait facile en gros de supprimer le papier et de faire comme s'il n'avait jamais existé. Dans le journal par exemple, quand on a fait une erreur, on revient dans un deuxième temps pour dire ‹Voilà, on a fait une erreur›, on explique le comment, le pourquoi et puis voilà. En fonction du deepfake c'est quelque chose qu'on va pousser plus loin avec le service juridique pour voir si on pousse jusqu'à des poursuites pénales, en fonction de la gravité d'effet, mais ça, après, ça va être du cas par cas.»*[246]

Allerdings erwähnen viele auch, das Dilemma zwischen Informieren und den Fehler nicht zu stark zu streuen. «Donc c'est aussi bien choisir quand réagir, pour ne pas renforcer la polémique ou lui donner plus d'écho en quelque sorte, ‹don't feed the Troll› comme on dit.»[247] Nichtstun scheint aber trotzdem keine Option zu sein. «Wenn wir das nicht korrigieren, dann wird das einer unserer Konkurrenten machen. Von daher ist das Bestandteil von unserer Arbeit.»[248] Deshalb sei es umso wichtiger, dass es gar nicht erst zu dieser Situation kommt. Viele verweisen auf das Problem, dass Informationen nur schwierig aus dem Internet zu entfernen sind. «Also für den Fall, dass wir ein Deepfake-Video auf

---

[244] Interview 3XC231122.

[245] Interview CDW161122.

[246] Interview HFK021122.

[247] Interview JOP071222.

[248] Interview TFX110123.

TikTok rauslassen und das wird dort irgendwie verbreitet: Nein, vergiss es. Das kannst du nicht mehr einholen. Da kannst du höchstens auf all den Kanälen, auf welchen du es rausgelassen hast, die Korrektur publizieren.»[249] Eine andere interviewte Person gibt zu bedenken, dass man mit einer Richtigstellung nur einen Teil des Publikums erreichen kann und natürlich die Skeptiker der Korrektur nicht glauben werden. «Wenn eine grosse Menge an Deepfakes verbreitet wurde und danach die Info kommt, dass es sich um eine Falschmeldung handelte, interessiert das gar nicht mehr. Man glaubt das schon und die Richtigstellung rückt irgendwie in den Hintergrund.»[250]

#### 5.3.2.5. Journalistinnen und Journalisten als Betroffene

*Fälle im Schweizer Journalismus:* Den Interviewten war kein Fall in der Schweiz bekannt, bei dem Deepfakes zur Diskreditierung von Journalistinnen und Journalisten verwendet wurden, so wie dies bereits in anderen Ländern vorkam (siehe z.B. Ayyub, 2018). «Aber ich habe jetzt nicht das Gefühl, das Leute in der Schweiz Angst vor solchen Kampagnen oder körperlichen Übergriffen haben.»[251] Dies hänge vielleicht auch damit zusammen, dass in der Schweiz verglichen mit dem Ausland die Reichweite von Persönlichkeiten im Journalismus eher limitiert sei. «En Suisse, les journalistes sont beaucoup moins médiatisés, beaucoup moins popularisés qu'en France ou aux États Unis. Ça ne change pas le fait qu'il y ait des figures très porteuses. Quand même un peu moins qu'à l'étranger.»[252] Dies lässt vielleicht das Risiko auch etwas vergessen, wobei manche durch das Interview auch angeregt wurden, über die Möglichkeit eines Angriffs mit Deepfakes nachzudenken. «On a tendance à peut-être pas vraiment y penser. Mais cela me sensibilise en quelque sorte à la possibilité d'un tel scénario.»[253]

Denn angegriffen werden Journalistinnen und Journalisten auch hierzulande, «aber es ist in der Schweiz immer noch sehr zivilisiert».[254] Hass und Attacken ohne Deepfakes sind häufig: «Hass im Internet ist sicher auch bei uns ein The-

[249] Interview SDE271022.

[250] Interview 3XC231122.

[251] Interview SDE271022.

[252] Interview HFK021122.

[253] Interview PAG171122.

[254] Interview SDE271022.

ma. Da gibt es auch diverse Workshops, wie man damit umgehen soll.»[255] Gerade bei investigativen Arbeiten exponieren sich Journalistinnen und Journalisten. «Ich habe ein Jahr über den sechsten Januar [Erstürmung des US-Kapitols 2021] recherchiert und mir wurde vorgeworfen, dass ich ein russischer Agent, ein Antifa und was weiss ich alles sei. Aber ich bin ziemlich hart im Nehmen. Ich habe die Feuertaufe schon hinter mir, nichts überrascht mich.»[256] Deshalb halten einige Interviewte eine Diskreditierung mit Deepfakes künftig durchaus für möglich. «Das ist natürlich ein ganz anderes Kaliber, also wenn man plötzlich in Deepfakes auftaucht und dann sagt, das stimme nicht, aber es glauben einem nicht alle, besonders in gewissen Kreisen.»[257] Die Interviewten wiesen aber darauf hin, dass diese Gefahr sie nicht an der Arbeit hindere. «Schlussendlich habe ich als Journalist auch die Möglichkeit, meine Plattform zu nutzen, um das zu demaskieren.»[258]

*Spezifische Strategien, Notfallpläne, Abteilungen:* Für den Fall, dass Journalistinnen und Journalisten selbst durch Deepfakes diskreditiert würden, sind den Interviewten keine spezifischen Notfallpläne bekannt. Oft seien Angriffe «eine Ad-hoc-Situation»[259] und es würden die bisherigen Strategien ergriffen, die für Attacken und Bedrohungen auf Journalisten gelten. «Rapidement communiquer pour donner notre point de vue, pour se positionner et au besoin déposer plainte soit le journaliste le fait en son nom et est soutenu par l'entreprise, soit l'entreprise si c'est une contrefaçon de la marque en quelque sorte.»[260] Um schnell zu reagieren, können dann die technischen Mittel und vielen Kommunikationskanäle wiederum eine Chance sein. «Das Opfer hat in der Regel die Möglichkeit, innerhalb von Sekunden die ganze Welt zu erreichen und zu sagen, dass es sich um ein Fake handelt. Die Technologie und die weltweite Sendemöglichkeit ist nicht nur ein Fluch, weil du gefaktes Zeug rauslassen kannst, sondern du kannst auch in Sekundenschnelle reagieren.»[261] Zusätzlich gebe es auch «diverse Dokumente, wie man sich bei Hate Speech verhalten soll, oder Stellen, an

[255] Interview 3XC231122.

[256] Interview TFX110123.

[257] Interview CDW161122.

[258] Interview HJK091122.

[259] Interview HJK091122.

[260] Interview JOP071222.

[261] Interview SDE271022.

die man sich wenden kann, wie zum Beispiel die Rechtsabteilung oder persönliche Ansprechpartner, die einen unterstützen».[262] Auch wenn immer wieder erwähnt wurde, dass gravierende Fälle in der Schweiz als eher unwahrscheinlich erachtet würden, denken einige Interviewte, dass die negative Stimmung gegen Journalistinnen und Journalisten in den letzten Jahren stärker wurde. «Das ist bei uns im Team ein Thema, worüber man spricht. Ja, wir nehmen das ernst. Ich glaube, früher wurde das oft zu wenig ernst genommen.»[263]

#### 5.3.2.6. Sensibilisierung der Nutzenden und Förderung von Medienkompetenz

*Sensibilisierung des Publikums als Win-win-Situation:* Ein Beitrag des Journalismus zur Sensibilisierung des Publikums für Deepfakes scheint auch vor dem Hintergrund der durchgeführten Bevölkerungsumfrage relevant, welche Defizite in der Identifizierung von, der Kenntnis über und des Umgangs mit Deepfakes in der Bevölkerung aufgezeigt hat (siehe Kapitel 3.3.3). Viele sehen in der heutigen Situation bezüglich Desinformation und hinsichtlich Deepfakes auch eine Chance für die Medien, um Menschen dabei zu helfen, kritischer zu denken und Informationen zu hinterfragen.

> *«Ich finde, das ist sowohl eine Gefahr als auch eine Chance. Klar, wenn alles angezweifelt wird oder für falsch gehalten wird, Stichwort QAnon oder irgendwelche Verschwörungstheorien, dann wird es schwer. Aber grundsätzlich finde ich ein gefördertes, kritisches Bewusstsein etwas Positives. Aber dort wäre ja eigentlich der Auftrag der Medien, dass sie eben mit der Verifikation von Informationen auch eine Einordnung vornehmen. Das ist etwas, wofür sicher ein Bedürfnis besteht, wo es eben auch gerade in unserer Verantwortung ist, dass wir richtige Fakten liefern.»*[264]

[262] Interview 3XC231122.

[263] Interview CDW161122.

[264] Interview HJK091122.

Dabei werden Deepfakes aufgrund ihrer Komplexität als «die Königsklasse der Einordnung»[265] bezeichnet. Für alle Interviewten ist es wichtig, dass in der Gesellschaft das Bewusstsein für Deepfakes vorhanden ist oder gefördert wird.

> *«Auch hier denke ich, dass es eine Frage des Bewusstseins ist. Auch auf die ersten gefälschten Steintafeln und Fotografien und Audiobeiträge und sehr wahrscheinlich auch auf die ersten Deepfakes ist man reingefallen. Bei sensationellen Videos auf YouTube Videos ist einer der ersten drei Kommentare immer ‹Das ist fake!›.»*[266]

Die Sensibilisierung des Publikums für Desinformation und Deepfakes stellt aus Sicht der Interviewten eine gute Möglichkeit dar, um das Publikum zu binden und für Journalismus zu gewinnen. «On essaye de faire ce travail, d'expliquer notre métier dans nos émissions, quand c'est nécessaire.»[267] Manche sprechen dabei sogar vom Aufbau oder Erhalt einer eigenen Community. «On essaye de construire une communauté autour de notre rédaction. Donc par exemple d'ouvrir, d'être transparent en ouvrant par exemple la rédaction à nos lecteurs qui peuvent parfois participer soit en ligne, soit physiquement à des séances de rédaction; qui peuvent nous poser des questions.»[268] Gerade in Zeiten des Vertrauensverlusts gegenüber dem Journalismus geht es vielen der Interviewten auch darum, immer wieder den Mehrwert journalistischer Arbeit gegenüber anderen Informationsquellen aufzuzeigen. «Journalismus kostet Zeit und Geld. Es ist wichtig, dass auch das Publikum realisiert, dass hinter qualitativer Berichterstattung ein grosser Aufwand steht, der bezahlt werden muss. Wir müssen die Gratiskultur im Journalismus überwinden.»[269]

*Sensibilisierung durch Berichterstattung:* Die Interviewten verwiesen auf verschiedene Möglichkeiten, wie das Publikum für Deepfakes sensibilisiert werden könnte. Auch der Journalismus könne einen Beitrag leisten, um die Medienkompetenz im Bereich der Informations- und Quellenkritik zu stärken:

---

[265] Interview QWE161122.

[266] Interview SDE271022.

[267] Interview JOP071222.

[268] Interview PAG171122.

[269] Interview CDW161122.

*«Im Sinne eines Servicecharakters können wir Geschichten schreiben und sagen: ‹Das ist ein Deepfake. Anhand von dem, dem und dem könnt ihr als Leser erkennen, dass es ein Deepfake ist. Auf das müsst ihr Acht geben, wenn ihr potenziell mit solchen Sachen konfrontiert seid.› Also dort sehe ich auch den Auftrag der Medien, die Gesellschaft zu sensibilisieren, dass halt solche Sachen vorkommen. Es begann mit gefälschten Bildern. Irgendwann dachte man sich, was Video ist, ist real. Aber jetzt ist nicht einmal mehr Video oder Live unter Umständen real.»*[270]

Die Sensibilisierung des Publikums müsse dabei aber zur Aktualität im journalistischen Alltag passen. «Je pense que ça pourrait se faire si c'est nécessaire. Encore une fois, c'est quelque chose qu'on fait quand on le juge utile. C'est à dire quand on juge qu'il y a un effet. Je pense qu'on a un travail à faire là-dessus, mais qu'on doit le faire quand ça a une importance.»[271] Da Desinformation derzeit ein relevantes Thema sei und sich nicht nur auf Deepfakes beziehe, gebe es immer wieder Berichte, die in Bezug zu Deepfakes stehen. «On fait régulièrement des articles sur d'autres choses qui sont un petit peu liées, mais en fait on est dépendants de nos sources et des nouvelles, et de la recherche aussi qui est faite dans ce domaine.»[272] Redaktionen thematisieren auch, wie sie selbst Fakten prüfen: «Was wir auch relativ häufig machen, sind allgemein Faktencheck-Beiträge, wo Desinformationen aufgenommen werden und breit gezeigt wird, wie man das entlarvt und wie das geprüft wird.»[273] Das komme auch gut beim Publikum an und die Journalistinnen und Journalisten könnten dabei von ihrem Wissen weitergeben. «Ich möchte eigentlich den Leser an die Hand nehmen und sagen: ‹Komm, ich zeige dir mal, wie ich das gemacht habe.›.»[274]

*Sensibilisierung durch weitere Aktivitäten:* Nicht nur in der Berichterstattung, auch über weitere Aktivitäten kann das Publikum für Deepfakes sensibilisiert werden. Ein Interviewpartner ist dezidiert der Meinung, der Journalismus müsse mehr Öffentlichkeitsarbeit leisten:

---

[270] Interview HJK091122.

[271] Interview JOP071222.

[272] Interview PAG171122.

[273] Interview 3XC231122.

[274] Interview QWE161122.

*«Ich habe neulich einen Vortrag vor mehreren hundert Leuten gehalten und da habe ich über meine Arbeit gesprochen. Ich habe gezeigt, wie wir Informationen verifizieren, wie wir unterscheiden, was ein Deepfake ist und was nicht. Diese Öffentlichkeitsarbeit ist sehr wichtig für Medien, damit die Leser uns vertrauen. Denn die haben keine Ahnung, was dahintersteckt. Ich will, dass die Leute mir vertrauen. Aber wir müssen ihnen zeigen, welche Methoden wir benutzen, wie wir die Technologie anwenden.»*[275]

Medien sind teilweise auch an Veranstaltungen mit eigenen Ständen vor Ort. «Wir hatten zum Beispiel letztes Jahr einen Stand am Digital Festival in Zürich, einen Stand zum Thema Fake News. Und da konnte man selbst einen Avatar mit einem Audio bespielen. Da bekamen wir sehr viel gutes Feedback.»[276]

Zur Sensibilisierung arbeiten einige Medien auch mit Schulen zusammen. «Je donne des cours dans des écoles pour sensibiliser les jeunes à ces deepfakes, et justement à comment être sûrs de l'information qu'ils voient, que ce soit une vraie information. Ça, ce sont des petites choses que les journalistes et les entreprises font pour plutôt sensibiliser les jeunes à ces problématiques.»[277]

Neben Vorträgen und Workshops in Schulen bieten viele Medienorganisationen auch Führungen und Projekte in Zusammenarbeit mit Schulen an. Der Nutzen sei dabei sowohl für die Schülerinnen und Schulen wie auch für die Medienorganisationen gross.

*«Im letzten Jahr hatten wir ein Projekt namens ‹Youth Lab›. Dabei handelt es sich um eine ausgewählte Gruppe von 25–30 Jugendlichen, die jede Woche zu uns kamen und dann behandelten wir jede Woche ein neues journalistisches Thema. Dabei ging es einerseits darum, herauszufinden, was die Jugendlichen interessiert, wie sie die Inhalte wollen, um so auch die junge Leserschaft bedienen zu können. Andererseits wollten wir ihnen auch etwas mit auf den Weg geben, das sie im Alltag nutzen können und da war das Thema Deepfake ein Teil davon.»*[278]

---

[275] Interview TFX110123.

[276] Interview 3XC231122.

[277] Interview HFK021122.

[278] Interview HJK091122.

Einige Interviewte meinen hingegen, das journalistische Angebot müsse für sich selbst sprechen. «Mit einem alters- und kanalgerechten Angebot kann man zeigen: nach diesen und diesen Kriterien ist es gemacht, das sind die Quellen, es ist nachvollziehbar, der hat es geschrieben, du kannst kommentieren, wir korrigieren, wenn etwas falsch ist.»[279]

*Skepsis gegenüber Sensibilisierung:* Trotzdem sehen viele auch Schwierigkeiten bei der Sensibilisierung des Publikums. Medienskeptische Personen könnten mit journalistischen Medien teilweise gar nicht erreicht werden und würden eher Informationen auf sozialen Netzwerken vertrauen, «wo es keine Kontrollinstanzen gibt».[280] «Wenn die Leute einmal in diesem Kaninchenbau sind, kommen sie da nur sehr schwierig raus. Das haben wir auch bei Covid gesehen. Man vertritt eine Meinung und bleibt dabei.»[281]

In den Interviews wurde deshalb auch hinterfragt, ob Deepfakes von Medien überhaupt thematisiert werden sollen, da jegliche Sensibilisierungsarbeit das Risiko berge, dass sich das Misstrauen gegenüber Medien noch verstärke.

> *«Ich stelle mich auf den Standpunkt, dass es für nach journalistischen Kriterien produzierende Medienorganisationen schädlich ist, wenn sie sagen: ‹Hey, wir trainieren jetzt, wie man richtige News von Fake News unterscheidet.›. Weil ich stelle mich auf den Standpunkt, dass es keine Fake News gibt. Es gibt News und es gibt Propaganda. Und wenn du selbst – und das ist der Fehler, den viele Medienorganisationen machen – den Begriff Fake News und alternative Fakten übernimmst, dann suggerierst du quasi, den Medien ist nicht zu trauen. Ich erachte es nicht als die Aufgabe von journalistischen Medienorganisationen, die Leute dafür zu sensibilisieren, was richtig und was falsch ist. In den Schulen wird das gemacht. Aber, dass journalistische Medienorganisationen selbst hingehen und sagen: ‹Wir müssen unser Vertrauen wiederherstellen›, das würde ich nicht machen. Weil das haben wir nicht nötig. Der Fake-News-Vorwurf kommt von der rechtspopulistischen Seite und der stimmt nicht.»*[282]

---

[279] Interview SDE271022.

[280] Interview SDE271022.

[281] Interview TFX110123.

[282] Interview SDE271022.

## 5.4. Hauptbefunde

Das aktuelle Kapitel bietet detaillierte Einblicke in den Umgang mit Deepfakes in Redaktionen von Schweizer Medienorganisationen und in der Journalismusausbildung. Zum Zeitpunkt der empirischen Erhebung (Herbst/Winter 2022) waren Redaktionen noch eher selten mit Deepfakes konfrontiert. Die Seltenheit von Deepfakes speziell im Schweizer Kontext hatte entsprechend Einfluss auf die Thematisierung in den Ausbildungsorganisationen wie auch auf die Einschätzung durch Vertreter von Medienorganisationen. Mit der zunehmenden Verbreitung und Zugänglichkeit der hinter Deepfakes stehenden Technologie in der breiten Bevölkerung scheint es aber durchaus realistisch, dass in Zukunft mehr Fälle auftreten werden.

Das Ziel der Studie bestand erstens darin zu untersuchen, welche Strategien in Redaktionen zur Identifikation von Deepfakes angewendet werden und zu welchen konkreten Anpassungen von Arbeitsprozessen und Routinen dies aktuell und künftig führt (FF 4.1). In der journalistischen Praxis werden Deepfakes als technischer Sonderfall von Desinformation betrachtet, die es im Rahmen der regulären Faktenüberprüfung zu erkennen gelte. Aus Sicht der Interviewten hilft beim Umgang mit Deepfakes die Orientierung an grundlegenden journalistischen Normen und Standards, welche in der Ausbildung erlernt und in der Praxis in Erinnerung gerufen und angewendet werden sollten. Dazu gehört auch, dass Informationen aus sozialen Netzwerken immer durch andere Quellen verifiziert werden sollten. In den untersuchten Medienorganisationen gibt es allerdings ein Bedürfnis nach Personen oder spezialisierten Abteilungen, die für komplizierte Fälle der Verifikation von Videos zuständig sind. Einige Redaktionen verfügen bereits über solche Teams oder Personen, um Inhalte aus sozialen Netzwerken oder Videos zu prüfen (Open-Source-Intelligence-Expertinnen und -Experten, Faktenchecker) resp. sind daran, solche Stellen aufzubauen. Da die finanziellen Ressourcen im Journalismus knapp sind, dürfte die Einsetzung oder der Ausbau von internen Verifikationsteams nicht bei allen Medienorganisationen realistisch sein. Auch ist es fraglich, wie verlässlich technische Tools zur Erkennung von Deepfakes künftig noch sind.

Die Ergebnisse zeigen, dass in der journalistischen Praxis kaum ein Unterschied zum umfassenderen Problem der Desinformation gesehen wird. Die Interviewten befürchten, dass Deepfakes den Vertrauensverlust in Medien und das unbegründete Hinterfragen von Fakten durch einen Teil der Nutzerinnen und Nutzer noch weiter verstärken werden. Diese Problematik wird durch Medienangebote verschärft, welche sich nicht an journalistische Standards halten. Zusätzlich

zu Verifikationsprozessen und Faktenüberprüfung wurde von den interviewten Personen deshalb die Sensibilisierung des Publikums für Deepfakes und die Information über die Verifikationsprozesse von Medien als konstruktiver Weg gesehen, um das Vertrauen in die Medien zu fördern. Es scheint also nicht ausreichend, wenn nur aufseiten der Medien Informationen verifiziert werden und ein Bewusstsein für manipulierte Information vorhanden ist, sondern es braucht in der ganzen Gesellschaft ein Bewusstsein für manipulierte Information und für die Notwendigkeit von Quellenkritik, was auf die Relevanz der Medienkompetenzförderung verweist. Dabei scheinen zwar kurze Interventionen oder Warnhinweise wenig effektiv zu sein (siehe Kapitel 3.3.3), allerdings könnten Medien eine wichtige Rolle dabei spielen, die Bevölkerung langfristig für Deepfakes zu sensibilisieren. Auch in der Wirtschaft wird Potenzial darin gesehen, Deepfake-Technologien in der Medienpädagogik zur Förderung einer Bewusstseinsbildung und Sensibilität für Falschinformationen zu verwenden (siehe Kapitel 7.2).

Insgesamt werden Deepfakes von den interviewten Personen also überwiegend im Kontext von Desinformation betrachtet und als Risiko wahrgenommen. So betrachtet liegt eine Chance für den Journalismus darin, sich durch die erfolgreiche Erkennung von Deepfakes und anderen Manipulationsversuchen als vertrauenswürdige Angebote von anderen Quellen abzuheben. Dennoch wurde von einigen Interviewten auch das Potenzial der Technologie hinter Deepfakes gesehen, etwa für verdeckte Recherchen, die Personalisierung durch synthetische Moderatorinnen und Moderatoren (Avatare), die lippensynchrone Übersetzung von Videos oder die Verbesserung der Zugänglichkeit. Allerdings wird auch auf das Spannungsverhältnis zwischen der Erstellung synthetischer Inhalte durch Redaktionen und der Achtung journalistischer Grundwerte hingewiesen.

Zweitens interessierte in der vorliegenden Studie, wie die Herausforderungen, welche Deepfakes für den Journalismus mit sich bringen, in der Ausbildung von Medienschaffenden aktuell und in Zukunft thematisiert werden (FF 4.2). Ähnlich wie in der journalistischen Praxis werden Deepfakes auch in der Ausbildung als Teil des Problems der Desinformation behandelt. Hauptziele der Ausbildung in Bezug auf den Umgang mit Deepfakes sind, dass angehende Journalistinnen und Journalisten Informationen gemäss professioneller Normen zu verifizieren versuchen und Deepfakes sowie aktuelle technologische Möglichkeiten verstehen. Dabei werden in den Ausbildungsorganisationen Deepfakes mehrheitlich als Beispiele in bestehenden Kursen behandelt. Die Orientierung an grundlegenden journalistischen Normen und Standards werde durch die Komplexität der Identifizierung von Deepfakes wichtiger. Diese Normen werden in der Ausbildung erlernt und mit Deepfakes in Verbindung gebracht. Durch die Verbesserung der Techno-

logie und deren zunehmenden Verbreitung und Zugänglichkeit kann davon ausgegangen werden, dass Deepfakes als Thema und Kompetenzen zur Überprüfung komplexer Information in der Ausbildung in Zukunft wichtiger werden. Zudem scheint eine kontinuierliche Weiterbildung von Medienschaffenden in Bezug auf technische Entwicklungen und Verifikationsmöglichkeiten unabdingbar.

Drittens stand in dieser Studie im Fokus, wie Medienorganisationen auf Fälle vorbereitet sind, in denen Journalistinnen und Journalisten selbst von Deepfakes betroffen sind (FF 4.3). Bisher sind Mitarbeitende von Schweizer Redaktionen noch nicht selbst Opfer von Deepfakes geworden. Strategien im Umgang mit Attacken wurden bereits als Reaktion auf Desinformation diskutiert und etabliert. Deepfakes werden dabei zwar als zusätzliches Risiko gesehen, das sich aber in das Grundproblem von Desinformation und Misstrauen in den Journalismus eingliedert. Ähnlich wie in der Politik (siehe Kapitel 6.2) scheinen auch im Journalismus konkrete Schutzmassnahmen selten zu sein. Da von Journalistinnen und Journalisten im Bereich Fernsehen und Video mehr Bild- und Videomaterial veröffentlicht wird, wird bei diesen auch das Risiko höher eingeschätzt, von einem Deepfake betroffen zu sein.

Die vorliegende Teilstudie konzentrierte sich aufgrund beschränkter Mittel auf die grössten Medienorganisationen der Schweiz. Diese verfügen trotz der Finanzierungskrise des Journalismus im Vergleich mit kleineren Medien über eine bessere Ressourcenausstattung und haben damit auch mehr Möglichkeiten, um in technische Systeme und spezialisierte Teams zur Erkennung von Deepfakes zu investieren. Insofern ist es durchaus möglich, dass der Umgang mit Deepfakes bei kleineren (zumeist regionalen und lokalen) Medienorganisationen sich von der hier dargestellten Situation unterscheidet. Ebenso spielt die Nachrichtenagentur Keystone-SDA für die Zulieferung von Nachrichten und Bildern bei kleinen Medien eine noch wichtigere Rolle. Insofern wäre auch von Interesse, aufzuzeigen, wie Keystone-SDA die Chancen und Risiken von Deepfakes für den Journalismus einschätzt und welche Massnahmen zur Erkennung von Deepfakes getroffen werden. Aus den Interviews ging hervor, wie wichtig es für Journalistinnen und Journalisten ist, sich speziell bei Videomaterial auf Quellen von Partnerorganisationen (z.B. im Falle des Service public andere Unternehmenseinheiten innerhalb der SRG und andere europäische Rundfunkorganisationen) und Agenturen verlassen zu können. Bei Keystone-SDA wurde deshalb 2020 eine Stelle geschaffen mit der Aufgabe, Behauptungen in sozialen Netzwerken zu verifizieren (Interview mit Frau Gilbert in Hochstrasser 2023). Diese Stelle steht im Austausch mit anderen Nachrichtenagenturen wie der DPA (Deut-

sche Presse-Agentur) und der APA (Austria Presse Agentur) und den Techunternehmen, welche die Plattformen betreiben.

Schliesslich sollte nicht unterschätzt werden, dass sich die hinter Deepfakes stehende Technologie sehr schnell entwickelt. Entsprechend wichtig ist es, dass Schweizer Medienorganisationen sich der daraus ergebenden Herausforderungen und Möglichkeiten für den Journalismus bewusst sind. Künftige Studien bieten auch die Möglichkeit zu untersuchen, wie sich das Bewusstsein bezüglich der Chancen und Risiken von Deepfakes bei Journalistinnen und Journalisten sowie Medienorganisationen weiter verändert.

# 6. Deepfakes in der Politik

*Murat Karaboga, Greta Runge & Michael Friedewald*

Schon bald nach Veröffentlichung der ersten Deepfakes Ende 2017 wurden Deepfake-Technologien in der wissenschaftlichen (Chesney/Citron 2018) und öffentlichen Debatte (Bezmalinovic 2020) die Möglichkeit zugerechnet, politische Prozesse zu manipulieren.

Die internationale und nationale Politik stellt ein Feld dar, das schon immer mit Desinformation, Verschwörungsmythen und Unwahrheiten konfrontiert war. Lüge und Täuschung waren und sind Teil politischer Auseinandersetzungen. Gleichzeitig ist Politik über «die Herstellung und Durchsetzung kollektiv verbindlicher Entscheidungen» (Buchstein 2012: 18) definiert. Und für die Verabschiedung solcher Entscheidungen bedarf es auch Wahrheiten, die von den am öffentlichen Diskurs beteiligten Akteuren geteilt werden. Ausgehend davon ist es die Aufgabe von Politikerinnen und Politikern, mittels unterschiedlicher Bewertung von Ereignissen, Sachverhalten usw. (also durch divergierende Haltungen zu politischen Fragen) die Unterstützung der Bevölkerung zu gewinnen. Gerade die Differenzierung zwischen Tatsachen und Meinungen stellt sich allerdings regelmässig als herausfordernd dar: Was für die einen ein Fakt ist, ist für andere blosse Meinungsäusserung ohne faktischen Wahrheitskern (Steinbach 2017). So wahr diese grundlegende Herausforderung auf der Meinungsäusserungsfreiheit basierender Demokratien mit Blick auf die Diskussion vieler politischer Fragen ist, lässt sie sich in der Debatte um Deepfakes zunächst doch etwas einfacher beantworten: Ein Deepfake ist ein Medieninhalt, der Menschen dabei zeigt, wie sie etwas tun oder sagen, was sie so nie gesagt oder getan haben. Somit handelt es sich bei einem Deepfake grundsätzlich um einen offensichtlich wahrheitswidrigen Inhalt. Auch eine mittels Deepfakes dargestellte z.B. fiktive Naturkatastrophe kann durch unterschiedliche Interpretation nicht wahr werden. Allerdings sind auch Kontexte denkbar, in denen ein Deepfake, ausgehend von einer Tatsache zur Darstellung der kontextspezifischen Interpretation, einer Situation dienen kann, über die es durchaus unterschiedliche Meinungen geben kann. Beispielsweise können die Worte einer Politikerin oder eines Politikers auf unterschiedliche Weise gedeutet werden. Ein Deepfake, der das Produkt einer solchen Deutung ist, könnte basierend auf den zuvor getätigten Aussagen einer Person weiterführende, mitunter satirisch überhöhte, Aussagen beinhalten. Diese unwahren Aussagen könnten von einigen Menschen trotzdem als offensicht-

lich zum Gedankengut der abgebildeten Person passend interpretiert werden, während ein anderer Teil des Publikums diese als beleidigend oder diffamierend interpretieren mag. Es ist zu erwarten, dass solche Verwendungen ihren Weg vor die Gerichte finden und je nach Fall sowohl zivilrechtliche oder strafrechtliche Folgen für die Erschaffer entstehen als auch Freisprüche unter Verweis auf die Meinungsfreiheit erfolgen werden. Denn selbst unwahre, falsche oder irreführende Aussagen unterstehen bis zu einem gewissen Grad dem grundrechtlichen Schutz (vgl. hierzu die zugehörige rechtliche Diskussion Kapitel 4). Unsere Studie und das vorliegende Kapitel haben daher nicht den Anspruch oder das Ziel, derartige Meinungsverschiedenheiten auszutarieren. Diese gehören, abgesehen von den juristisch zu entscheidenden Fällen, in das Zentrum der öffentlichen Debatte und sollten von der Gesellschaft geführt werden.

Andererseits wird Deepfakes zugetraut, den politischen Diskurs dermassen zu verzerren, dass Politikerinnen und Politiker diffamiert, demokratische Institutionen und die Demokratie insgesamt beschädigt werden können (vgl. hierzu die nachfolgende Diskussion des Forschungsstands zu Deepfakes in der Politik). Vor dem Hintergrund dieser Debatte verfolgt das vorliegende Kapitel das Ziel, die gegenwärtige und künftige Rolle von Deepfakes in der Politik zu untersuchen. Hieraus werden wiederum Handlungsempfehlungen formuliert, wie allfällige Schäden aus gegen Politikerinnen und Politiker, politische Institutionen und das politische System gerichteten Deepfakes verhindert oder eingedämmt werden können.

Hierfür wird zunächst anhand einer Literaturanalyse der Stand der Forschung zu Deepfakes in der Politik aufgearbeitet (vgl. Kapitel 6.1). Zur Untersuchung der Wahrnehmung von Deepfakes in der Schweizer Politik führten wir eine Umfrage durch, in der Schweizer Parlamentarierinnen und Parlamentarier befragt wurden, deren Ergebnisse im nachfolgenden Kapitel vorgestellt werden (vgl. Kapitel 6.2). Aufbauend auf der Literaturanalyse und den empirischen Befragungsergebnissen werden anschliessend in Form einer Kurzzusammenfassung die Ergebnisse aus elf konkreten Szenarien vorgestellt, wie die Verwendung von Deepfakes politische Implikationen nach sich ziehen können (vgl. Kapitel 6.3).

## 6.1. Forschungsstand zu Deepfakes in der Politik

Die möglichen Auswirkungen von Deepfakes auf die Politik sind eingebettet in den digitalen Strukturwandel der Öffentlichkeit. Daher werden im folgenden Unterabschnitt (vgl. Kapitel 6.1.1) zunächst die Eckpunkte dieses Strukturwandels

näher beleuchtet. Im Anschluss werden die möglichen Auswirkungen von Deepfakes auf die Politik unter Einbeziehung exemplarischer realer Verwendungen von Deepfakes in der Politik diskutiert (vgl. Kapitel 6.1.2).

### 6.1.1. Der digitale Strukturwandel der Öffentlichkeit

Inwiefern ist der digitale Strukturwandel der Öffentlichkeit relevant im Hinblick auf politische Prozesse und demokratische Gesellschaften? Die Digitalisierung charakterisiert sich als soziotechnischer Prozess zunächst «durch tiefgreifende, vielfältige gesellschaftliche Transformationsprozesse» (Jarke 2018). Diese ordnen das Verhältnis zwischen privatem und öffentlichem Diskursraum neu, indem deren Grenzen verschoben werden. Was vor der Entstehung und Verbreitung der modernen Informations- und Kommunikationstechnologien noch im privaten Rahmen blieb, lässt sich mittels Social-Media-Plattformen in Sekundenschnelle in die gesamte Welt versenden. Der öffentliche Diskurs, der zuvor weitestgehend unidirektional strukturiert war, und damit auch die politische Öffentlichkeit, wurden im Ergebnis offener, interaktionsorientiert, multidirektional und nicht zuletzt komplexer und unübersichtlicher (Habermas 2022, S. 146 ff.). Soziale Netzwerke und Plattformen sind für politische Akteure zum unverzichtbaren Ort der Selbstinszenierung, der politischen Kommunikation und der digitalen Kampagnenführung avanciert (Jungherr 2022). Hier treffen sie auf ihre potenziellen Wähler und Bürgerinnen. Diese wiederum nutzen soziale Medien nicht nur für den persönlichen Austausch, sondern nutzen diese neuen Medien als zunehmend integralen Bestandteil ihrer Alltagskommunikation, indem sie dort Informationen jedweder Art suchen und teilen. Aus dem Verhalten der Menschen, die diese Technologien nutzen, ergeben sich neue gesellschaftliche Agenda-Setting-Dynamiken. Einzelpersonen, die vor dem digitalen Strukturwandel nur mit grosser Mühe und insb. auf Basis enormer Geldressourcen in der Lage waren, ihrer Stimme Gehör zu verschaffen, sind inzwischen, z.B. in Form von Influencern, zu wirkmächtigen Akteuren im politischen Diskurs geworden. Klassische Medien und traditionell mächtige Akteure aus Politik und Wirtschaft konkurrieren mit diesen *Content Creators* um die Aufmerksamkeit des Publikums (Bogner u.a. 2022). Dass viele Menschen ihre Informationen über diese neuen Plattformen beziehen, hat auch Folgen für die klassischen Medienorganisationen. Teils neben und teils an die Stelle dieser sind die neuen Social-Media-Plattformen getreten. Damit strukturieren Plattformen wie Facebook, X, Instagram, TikTok und viele weitere als neue mächtige Intermediäre den öffentlichen Diskurs mit (Jarren/Fischer 2022). Themen und Meinungen finden längst nicht mehr ausschliesslich über den Journalismus Eingang in öffentliche Debatten, sondern

es bilden sich zunehmend fragmentierte Teilöffentlichkeiten heraus, in denen Algorithmen und Bots Inhalte und Nachrichten strukturieren, deren ausschliessliche Rezeption zu selektiven Wahrnehmungen und zu einer Beeinflussung der freien Meinungs- und Willensbildung führen kann (Oertel u.a. 2022) (vgl. auch Kapitel 5).

Plattformen sind jedoch nicht nur Mittel der Interaktion und Kommunikation, sondern ihrer Logik nach selbst Ort der Datengenerierung und -auswertung geworden (Gentzel 2022). Es besteht eine massive Zunahme der Sammlung und Verarbeitung gesellschaftspolitisch relevanter Daten. Die datengenerierende Kommunikation und Transaktion im öffentlichen Raum führt zu «neuen Möglichkeiten der Beobachtung, der Vorhersage, aber auch der Manipulation von potenziell politisch relevantem Verhalten» (Borucki u.a. 2020). Automatisierte Systeme prägen verstärkt die Interaktion und Kommunikation zwischen Personen, Organisationen und Staaten. Der Diskurs um die gesellschaftlichen Auswirkungen der Digitalisierung wird durch KI-Technologien befeuert, welche die Verarbeitung und Analyse grosser Datenmengen (Big Data) implizieren. KI basiert auf handlungsleitenden Algorithmen, die inzwischen «zur allgegenwärtigen Infrastruktur der Gesellschaft selbst geworden» sind (Block/Dickel 2020: 112). KI kann zur präzisen Beobachtung, Überwachung, Filterung und Zensur von Nachrichten zwischen Nutzenden von Anwendungen und Plattformen eingesetzt werden (EDSB 2018). So liegt es nahe, dass die durch Daten dokumentierte Kommunikation und Information von Gesellschaften nicht nur für wirtschaftliche, sondern auch für politische Akteure interessant ist. Die Digitalisierung und die durch sie hervorgebrachten Technologien fordern die informationelle Selbstbestimmung und die Privatheit der Gesellschaftsmitglieder durch die kontinuierliche Zugänglichkeit zu personenbezogenen Daten und deren Verarbeitung heraus. Enorme Mengen an teils qualitativ hochwertigen personenbezogenen Daten(-banken) ermöglichen tiefe Einblicke in die Verhaltensweisen der Betroffenen und bieten somit eine umfassende Wissensgrundlage, auf der aufbauend mächtige Akteure in der Lage sind, Individuen und Gruppen zu manipulieren (Ritzi 2017). Die strukturellen Eigenschaften der Digitalisierung stehen in erheblicher Spannung zu klassischen territorialen Ordnungsvorstellungen und staatlichen Steuerungsmöglichkeiten, indem sich die Plattformen dem nationalstaatlichen Rechtszugriff entziehen (Borucki u.a. 2020). Damit wird nicht nur der nationalstaatliche Regulierungsspielraum eingeengt, sondern auf ambivalente Weise auch die individuelle Souveränität im digitalen Raum. Einerseits hat der fehlende nationalstaatliche Zugriff auf das Internet und auf Plattformen für viele Menschen, insb. in autoritären Staaten, Möglichkeiten der freien Entfaltung jenseits nationalstaatlicher Kontrolle geschaffen. Andererseits leiden Menschen gerade unter diesem

unzureichenden nationalstaatlichen Zugriff, wenn sie beispielsweise Opfer von Morddrohungen, Hassrede, Rufmordkampagnen, Einschüchterungsversuchen, Stalking usw. werden und sich nur unzureichend dagegen wehren können, weil die Plattformen die Kooperation oder gar die Einhaltung nationaler Gesetze ablehnen (Ritzi/Zierold 2019: 38; vgl. Kapitel 4.2).

Der digitale Strukturwandel der Öffentlichkeit fordert die Gesellschaft also besonders auf drei Weisen heraus, die auch für den Kontext Deepfakes und synthetische Medien relevant sind: 1. Es findet eine enorme Ausweitung der (zivil-) gesellschaftlichen Einflussnahme auf gesellschaftliche Diskurse statt. 2. Neben klassische Medienorganisationen treten Plattformen als machtvolle Informationsintermediäre, die den gesellschaftlichen Diskurs moderieren und steuern. 3. Das datengenerierende und datengetriebene Handeln aller am politischen Geschehen beteiligten Akteure (von den Bürgerinnen und Bürgern, über organisierte Akteure bis zu den Politikerinnen und Politikern selbst) schafft das Potenzial der Manipulation gesellschaftlicher Diskurse. Politische Akteure, die im Fokus dieses Kapitels stehen, sind mit Blick auf diese drei Entwicklungen also herausgefordert, auch im digitalen Raum selbstbestimmt handeln zu können und neue Arrangements zwischen individuellen und gesellschaftlichen Kommunikationsinteressen und dem politischen Steuerungsinteresse zu finden (Schallbruch 2018; Knorre et al. 2020).

### 6.1.2. Mögliche Implikationen von Deepfakes für die Politik

Nachdem das Thema Deepfakes von der Forschung aufgegriffen worden war, entstand eine Reihe von Publikationen, in denen Missbrauchsmöglichkeiten von Deepfakes für die Politik identifiziert wurden, darunter die Verwendung zur Diskreditierung von Politikerinnen und Politikern oder von demokratischen Institutionen, die Destabilisierung politischer Prozesse und die Beeinflussung von Wahlen, die Befeuerung inner- oder zwischenstaatlicher Konflikte und die Verwendung von Deepfakes zur Täuschung des Gegenübers, z.B. um Zugang zu geschützten Bereichen oder Informationen zu erhalten. Auch die reale Nutzung von Deepfakes zu politischen Zwecken hat inzwischen zugenommen. Gleichwohl gibt es in der wissenschaftlichen Literatur Nutzungsszenarien von Deepfakes, die so noch nicht in die Tat umgesetzt wurden. Im Folgenden werden die verschiedenen Nutzungsmöglichkeiten von Deepfakes in der Politik entlang der entsprechenden Literatur diskutiert und, wo vorhanden, auf reale Beispiele verwiesen. Das Unterkapitel hat somit zum Ziel, aufzuzeigen, wie Deepfakes schon

heute in der Politik eingesetzt werden und welche konkreten Herausforderungen, aber auch welche Chancen bestehen (Whittaker et al. 2020).

Der Einsatz von Deepfakes im Kontext politischer Prozesse fokussiert sich bislang auf Politikerinnen und Politiker, die unfreiwillig in synthetischen Medieninhalten abgebildet werden. Öffentlichkeitswirksame frühe Beispiele sind der Donald-Trump-Deepfake einer belgischen Partei, Nancy Pelosi, als vermeintlich «betrunkene» Sprecherin, oder Deepfakes rund um den ukrainischen Präsidenten Volodimir Selenskyj. Alle Beispiele eint die Absicht, Einfluss auf politische Prozesse zu nehmen. Hierbei bestehen jedoch unterschiedliche Strategien und Zielsetzungen, wie bei näherer Betrachtung dieser und weiterer Beispiele noch deutlich wird. Ausserdem können kaskadierende Effekte und Ziele auftreten bzw. vorherrschen. Ein Deepfake, der die Diskreditierung einer Person zum Ziel hat, kann zugleich die Diskreditierung ihrer Partei und auch die Beeinflussung einer Wahl beabsichtigen (van Huijstee u.a. 2021: 34 ff.). Zudem muss für das Eintreten eines schädigenden Effekts eines Deepfakes keine Schädigungsabsicht vorliegen. Angesichts der zunehmenden technologischen Fähigkeiten zur Deepfake-Produktion ist auch damit zu rechnen, dass mehr Menschen solche Inhalte produzieren und veröffentlichen werden, sodass mit einer Zunahme unbeabsichtigter Effekte zu rechnen ist.

Der bislang am weitesten verbreitete Typus von politischen Deepfakes ist die Diskreditierung einer einzelnen Politikerin oder eines Politikers. Ein Beispiel dafür ist das *Cheapfake* der früheren Sprecherin des Repräsentantenhauses der Vereinigten Staaten, Nancy Pelosi, als vermeintlich «betrunkene» Sprecherin auf einer Pressekonferenz. Das Video stellte sich zwar recht schnell als einfache Manipulation eines echten Videos heraus, das lediglich in langsamerer Geschwindigkeit veröffentlicht worden war. Der dadurch erreichte Effekt, dass Pelosi betrunken klang, wurde jedoch vom politischen Gegner aufgegriffen und dazu genutzt, die Politikerin zu kritisieren. Führende republikanische Politikerinnen und Politiker verbreiteten das Video und viele Social-Media-Nutzende taten es ihnen gleich, bis andere Medien das Video als Manipulation entlarvten (Harwell 2019).

Der Einsatz von Deepfakes im Wahlkampf gewann insbesondere im Verlauf der zweiten Präsidentschaftskampagne von Donald Trump an Bedeutung und wurde Teil der politischen Kommunikation (Meckel/Steinacker 2021). Deepfakes als Mittel für eine gezielte Täuschungsabsicht können Kandidatinnen und Kandidaten bei Handlungen oder der Tätigung von Äusserungen zeigen, die sie in Wirklichkeit nie gemacht haben, um sie in ihren Ansehen zu schädigen und so den Wahlausgang in eine gewünschte Richtung zu lenken. Jedoch ist nicht nur

die Schädigung einer einzelnen Person durch das Deepfake-Video denkbar – auch das Ansehen und im Ergebnis die Wahlchancen ihrer Partei könnten beeinflusst werden (van Huijstee u.a. 2021). Im März 2023 wurde im Rahmen des Wahlkampfes um das Amt des Bürgermeisters von Chicago ein Deepfake des Politikers und Bürgermeisterkandidaten Paul Vallas über Twitter veröffentlicht. Der Account *Chicago Lakefront News*, ein Nachrichtenmedium, welches es tatsächlich gar nicht gibt, postete den Deepfake. Das Video wurde schnell gelöscht und das Konto bei Twitter gesperrt. Ob dem Kandidaten durch den Voice-Over-Deepfake Schaden entstand, ist unklar. Weil eine Möglichkeit zur Überprüfung der Glaubwürdigkeit eines Inhalts die Überprüfung der Vertrauenswürdigkeit der Quelle ist, sollte mit der Schaffung des fiktiven Nachrichtenmediums *Chicago Lakefront News* eine vermeintlich vertrauenswürdige Quelle geschaffen werden. Auf diese Weise sollten Betrachter, die lediglich den Namen der Quelle lesen, reingelegt werden, sodass nur jene Personen, die nähere Details zur Quelle recherchieren, merken würden, dass es sich um eine gefälschte Nachrichtenseite handelt (Lewis 2023).

Durch die Verbreitung irreführender Inhalte im Rahmen von Desinformationskampagnen können auch diplomatische, internationale Beziehungen beeinflusst und die Innen- und Aussenpolitik destabilisiert und geschädigt werden. Konflikte und Unsicherheiten zwischen und innerhalb von Staaten können durch die Verbreitung von Deepfakes hervorgerufen oder verschärft werden, etwa indem Politikerinnen und Politikern Worte mit Brisanz in den Mund gelegt werden (Metzger/Schneider 2022). Vor Beginn des russischen Angriffskrieges gegen die Ukraine war bekannt geworden, dass Russland erwog, mittels einer Operation unter falscher Flagge einen Angriff auf das eigene Territorium der Ukraine anzulasten, um eine Rechtfertigung für den Angriffskrieg zu inszenieren und damit die internationale Politik zu beeinflussen. Die Pläne wurden bereits vorab vom US-Geheimdienst publik gemacht, weshalb es nie dazu kam (Sganga 2022). Angesichts des Aufwands solcher False-Flag-Operationen und der zunehmenden Möglichkeiten, mittels KI künftig nicht nur Gesichter, sondern auch vollständige Videosequenzen zu synthetisieren, ist damit zu rechnen, dass die Nutzung von Deepfake-Technologien zu derartigen Zwecken zunehmen wird. Ein Versuch der politischen Destabilisierung der Innenpolitik fand kurz nach Beginn des Krieges statt. Mittels eines im März 2022 seitens Russlands verbreiteten Deepfake-Videos des ukrainischen Präsidenten Volodimir Selenskyj, worin dieser seine Bevölkerung und Streitkräfte vermeintlich zur Kapitulation aufruft, wurde versucht, die ukrainische Bevölkerung zu verunsichern. Zur Steigerung der Effektivität des Videos hackten sich die Angreifer in die Social-Media-Kanäle des ukrainischen Fernsehsenders «Ukraine 24» und verbreiteten das Video darüber. Später wurde seitens des Senders ein Gegenvideo des echten ukrainischen

Präsidenten geteilt und YouTube, Facebook und Twitter nahmen den Deepfake von ihren Plattformen (Bolzern/Bucher 2022). Angesichts derartiger Nutzungen werden Deepfake-Technologien inzwischen als ein weiteres Element im Informationswettrüsten angesehen (van Huijstee u.a. 2021: 52).

Die Möglichkeiten der Manipulation des Meinungs- und Willensbildungsbildungsprozesses in der Bevölkerung durch Deepfake-Inhalte werden durch den im vorangegangenen Kapitel beschriebenen digitalen Strukturwandel der Öffentlichkeit gesteigert (Zimmermann/Kohring 2018). Die KI-getriebene Informationsverbreitung über Social Media, personalisierte Werbung und Deepfake-Inhalte wirken dabei ineinander und erhöhen die Gefahr der Manipulation des öffentlichen Diskurses (Heesen u.a. 2021). Dieses Ineinanderwirken wird im Folgenden näher beschrieben.

Die KI-getriebene Informationsverbreitung über Social Media basiert darauf, dass die Nutzenden durch Likes, Retweets, Sharing usw. zur Bewertung und Verbreitung von Inhalten beitragen. In der Vergangenheit basierte die Manipulation dieser Dynamik auf der Beschäftigung von Menschen, die mit den Aufgaben des Likens, Retweetens usw. befasst waren. Dutzende, sowohl autoritäre als auch demokratische, Staaten unterhalten Cybertruppen (häufig auch abschätzig als Troll-Farmen bezeichnet), mit denen sie die Beeinflussung der Bevölkerungsmeinung im eigenen Land oder in anderen Staaten praktizieren (Bradshaw/Howard 2017). Mit zunehmenden Möglichkeiten künstlicher Intelligenz werden diese Aufgaben allerdings von Menschen an Maschinen ausgelagert. Weil einfache Formen des automatisierten Likens, Retweetens usw. durch die Plattformbetreibenden leicht erkannt werden können, werden KI-Technologien dazu genutzt, die Erkennungsbemühungen der Plattformen zu täuschen.

Deepfake-Bilder sowie -videos und KI-generierte Texte eignen sich auch dazu, Social Bots als Menschen zu tarnen. In der einfachsten Variante lassen sich durch die Techniken des *Gesichtsmorphings* und der *Gesichtsgenerierung* Profilbilder vermeintlicher Nutzerinnen und Nutzer erzeugen, die so kein zweites Mal im Internet auftauchen und sowohl von Detektoren als auch vom menschlichen Auge kaum als unecht zu erkennen sind (Whittaker et al. 2020). Mittels moderner Text-Generatoren kann unterschiedlichen Profilen eine unterschiedliche, glaubwürdig wirkende Persönlichkeit gegeben werden, sodass die Bots in Social-Media-Postings authentisch in ihrer jeweiligen Rolle agieren. Damit könnten sich Social Bots unauffällig unter die restlichen Nutzenden mischen, bis ein Einsatzbefehl z.B. zum Zwecke einer Desinformations- oder Rufmordkampagne kommt (van Huijstee u.a. 2021: 51; Bateman 2020: 7). Nach Start einer solchen Kampagne können unter Rückgriff auf KI-basierte Social Bots schnell

irreführende Inhalte auf Plattformen geteilt werden, bei denen es so wirkt, als kämen sie von einem realen User oder einer Masse von Usern (Deutscher Bundestag 2020: 465). Deepfakes können zwar auch von menschlichen Usern absichtlich oder unabsichtlich weiterverbreitet werden und schnell viele Menschen erreichen. Dies vollständig zu steuern, kann für die Deepfake-Erschaffer allerdings schwierig sein. Der Rückgriff auf Social Bots kann die Erreichung eines grossen Publikums mit grösserer Gewissheit gewährleisten. Durch das Teilen eines Deepfake-Inhalts durch eine Masse an Social Bots kann der Inhalt z.B. in den Trending Content-Bereich einer Social-Media-Plattform gehievt werden, um schnell sehr viele Menschen zu erreichen (Guglielmi 2020; Rossetti/Zaman 2023).

Irreführende Mobilisierung im Internet kann auch seitens PR-Agenturen oder Organisationen (ein bekanntes Beispiel hierfür ist die russische *Internet Research Academy*) (Pallaske 2022) entspringen. Werden von diesen Akteuren und Organisationen Meinungen konstruiert und massenhaft verbreitet, ist von *Astroturfing* die Rede. Astroturfing beschreibt die Imitation einer Graswurzelbewegung, indem so getan wird, als stecke eine soziale Bewegung hinter Meinungen, die z.B. auf sozialen Netzwerken kursieren oder in öffentlichen Konsultationen (Mahbub u.a. 2019). Nutzen Astroturfer digitale Werkzeuge, ist von digitalem Astroturfing die Rede. Insbesondere mittels Algorithmeneinsatzes, um z.B. synthetische Inhalte zu generieren, ist eine Flut an meinungsgefärbten Tweets und Posts denkbar. Darüber hinaus könnten sich auch Reaktionen und Kommentare auf *Astroturf*-Meldungen automatisieren lassen, was negative Effekte wie Täuschungen und kampagnenbezogene Manipulation der User, beispielsweise in Bezug auf Informationen im Kontext von Wahlkämpfen, nach sich ziehen könnte (Kalpokas/Kalpokiene 2022). Werden unter einem Post Astroturf-Kommentare einer vermeintlichen Öffentlichkeit generiert, könnten Politikprozesse im Bereich des Agenda-Settings, Problemformulierung und Regelsetzung manipuliert und destabilisiert werden (Bateman 2020: 24). Bislang sind die realen Auswirkungen von Astroturfing umstritten, dennoch ist das Phänomen ein wachsendes regulatorisches Problem, das Potenziale birgt, Vertrauen der Öffentlichkeit in Institutionen, Gesetzgebungsprozesse und freie Wahlen zu untergraben (ebd.).

Auf Basis von Social-Media-Nutzungsdaten können Deepfakes, die z.B. Politikerinnen und Politiker abbilden, gezielt auf bestimmte, beeinflussbare Profile ausgerichtet und ausgespielt werden, um das (Wahl-)Verhalten zu beeinflussen. Die aus der Werbe- und Marketingbranche stammende Methode, speziell zugeschnittene Inhalte wie Deepfakes zu veröffentlichen, die eine bestimmte als beeinflussbar geltende Zielgruppe ansprechen, nennt sich Microtargeting. Die Verwendung

der Methode für politische Zwecke ist politisches Microtargeting (van Huijstee u.a. 2021: 24). Hierfür werden aus vorliegenden Nutzerdaten Persönlichkeitsprofile erstellt, um die Beeinflussung der Wahlentscheidung der Bürgerinnen und Bürger mittels äusserst zielgruppenspezifischer Werbung zu erreichen. Fraglich ist bislang, inwieweit sich diese Form der personalisierten Ansprache (insb. auch in Form gezielt verbreiteter Deepfakes) auf die Einstellungen der Bevölkerung zu dem politischen Akteur (oder auch der Partei) tatsächlich auswirken (Dobber u.a. 2021). Möglich sei jedenfalls, dass mittels politischem Microtargeting das politische Mobilisierungsverhalten beeinflusst werden könnte, indem mehr oder weniger emotionalisierende oder unterhaltende Inhalte an unterschiedliche Gruppen ausgesendet werden. Wenn die eigene Kernwählerschaft im Fokus der politischen Kommunikation steht, geht es um deren Mobilisierung im Interesse der eigenen Partei. Darüber hinaus können (insb. wenn konkurrierende politische Kandidaten gleichauf liegen) auch Unentschlossene oder Wechselwähler angesprochen werden, um ihre Wahlpräferenz zu beeinflussen (Magnani 2022). Schliesslich kann auch die Beeinflussung der Wählerschaft der Gegenseite das Ziel sein: So seien bei der US-amerikanischen Präsidentschaftswahl im Jahr 2020 Wählern, die den Demokraten zugeneigt sind, eher unterhaltende Inhalte ausgesendet worden, um ihren Mobilisierungswillen zu reduzieren. Währenddessen seien republikanische Wähler mittels polarisierender Inhalte zu öffentlichem Protest angestachelt worden (Cirone/Hobbs 2023).

Neben der absichtsvoll-missbräuchlichen Nutzung können Deepfakes auch ohne Absicht Schäden anrichten. 2018 wurde ein Deepfake seitens einer belgischen Partei (Burchard 2018) mit dem Ziel verbreitet, eine öffentliche Debatte zur Bedeutsamkeit des Klimaschutzes anzustossen. Die Kommunikationsmassnahme geriet allerdings in die Kritik, da das Deepfake nicht als solches gekennzeichnet und damit potenziell irreführend war. Im Frühjahr 2023 geriet Amnesty International nach der Verwendung KI-generierter Bilder zur Schaffung von Aufmerksamkeit für die Menschenrechtsverletzungen der kolumbianischen Polizei in die Kritik. Einer der zentralen Kritikpunkte war der Vorwurf, dass Amnesty International durch die Verbreitung KI-generierter Bilder an der allgemeinen Informationsverschmutzung und dem Vertrauensverlust in Informationen und Nachrichtenmedien mitwirke (Taylor 2023). In den Bereich der unabsichtsvollen Schädigung handelt es sich auch bei Misinformation. Schliesslich teilen viele Social-Media-Nutzende Desinformation, ohne sich darüber im Klaren zu sein, dass es sich dabei um Desinformation handelt.

Aufgrund der technologischen Fortschritte und niederschwelligen Zugänglichkeit von Deepfake-Software besteht künftig verstärkt die Möglichkeit, dass demokra-

tische Prozesse durch die Technologie negativ beeinflusst werden und heute unabsehbare Folgen auf die politische Meinungs- und Willensbildung resultieren (Heesen et al., 2021, S. 28). Digitale, durch Deepfakes realisierte Desinformation kann auch einen negativen Einfluss auf die öffentliche Wahrnehmung von Journalismus mit sich bringen. Die Risiken von Deepfakes als Werkzeug für Desinformation lassen sich nach Bovenschulte auf zwei unterschiedlichen Ebenen und Wirkungsweisen fassen: «Während gezielte Deepfakes insbesondere die Integrität von Personen genauso wie Institutionen infrage stellen, beträfe das steigende Misstrauen in Medienhalte die Gesellschaft und Demokratie als solche» (Bovenschulte, 2019, S. 4). Diese allgemeine Vertrauenserosion wiederum kann zwei Ausprägungen haben: wachsendes Misstrauen in die Glaubwürdigkeit realer Inhalte einerseits und das «Dividende des Lügners»-Phänomen andererseits (vgl. auch Kapitel 5). Welche Folgen die allgemeine Vertrauenserosion in erster Hinsicht nach sich ziehen kann, demonstrierte ein gescheiterter Militärputsch in Gabun. Nachdem Präsident Ali Bongo zuvor monatelang nicht mehr in der Öffentlichkeit aufgetreten war, entfachte eine im Dezember 2018 veröffentlichte Videobotschaft des Präsidenten eine öffentliche Debatte über die Authentizität des Videos. Einige Militärs nutzten diese öffentliche Irritation als Vorwand für einen Putschversuch, der später scheiterte (Cahlan 2020). Sollte die Verbreitung von Deepfakes das Vertrauen in Medieninhalte grundsätzlich untergraben, kann es auch zum umgekehrten Phänomen kommen: zum sog. «Dividende des Lügners»-Phänomen, sodass echten Ton- und Videoaufnahmen, die z.B. Politikerinnen und Politiker während korrupter Handlungen zeigen, nicht geglaubt wird (Fallis 2021). Selbst besonders skandalöse (und deshalb vielleicht unglaubwürdig) klingende oder aussehende Inhalte, die in schlechter Qualität vorliegen (beispielsweise die Ibiza-Affäre, welche die österreichische Politik erschütterte), könnten dann einfacher geleugnet werden. Beispielsweise brachten Verschwörungsideologen nach der Ermordung von George Floyd das Gerücht in Umlauf, dass alles eine schauspielerische Inszenierung gewesen sei (Alba 2020). Auf ähnliche Weise könnten auch der Wahrheitsgehalt beliebiger politischer Skandale geleugnet werden.

Neben visuellen Botschaften können durch Deepfakes auch KI-generierte Texte wie Falschnachrichten («Fake News») oder erfundene Zitate erzeugt und in Umlauf gebracht werden, die wie ein journalistischer Text wirken und auch die Schreibweise von Menschen imitieren können (EPTA 2023; Diresta 2020). Die damit provozierte Desinformation der Bevölkerung könnte zu einem Vertrauensverlust und zu Demokratieverdrossenheit führen. Die Anleitung der Bevölkerung zum Ausbau ihrer digitalen Kompetenzen kann den desinformativen Effekt zwar reduzieren. In einem gesellschaftlichen Klima allgegenwärtiger synthetischer

und manipulierter Medien- bzw. Nachrichteninhalte und in Kombination mit Phänomenen wie der Nachrichtenübermüdung (AP 2008) könnte das Setzen auf individuelle Medienkompetenz jedoch auch den gegenteiligen Effekt mit sich bringen und zu einem generellen Misstrauen gegenüber jeglichen Medien- und Nachrichteninhalten führen und der allgemeinen Vertrauenserosion Vorschub leisten.

Eine Vielzahl unterschiedlicher Akteure kann Desinformation betreiben. Wie die diskutierten Fälle realer Deepfake-Verwendungen zeigen, handelt es sich dabei zunächst vor allem um Akteure aus der Politik selbst, die solche Inhalte weiterverbreiten. Neben Politikerinnen und Politikern sind es ausserdem häufig politisch interessierte Menschen, die entsprechende Inhalte weiterverbreiten. Wie die internationalen Fälle zeigen, kann es sich dabei auch um staatlich gesteuerte oder staatlich unterstützte Akteure handeln. Denkbar ist auch, dass terroristische Gruppen Deepfakes zur Erreichung politischer Ziele nutzen. Angesichts der realen Verwendungen von Deepfakes zu Zwecken der Wirtschaftskriminalität (vgl. Kapitel 7) könnten kriminelle Vereinigungen auch politische Deepfakes zur Gewinnerzielung, z.B. durch Erpressung, verwenden. Schliesslich teilen viele Menschen Desinformation ohne Schädigungsabsicht (Misinformation), sodass ein schädigender Effekt auch unbeabsichtigterweise eintreten kann.

Die Deepfake-basierte Überwindung biometrischer Systeme, aber auch die Täuschung und Manipulation von Menschen mittels Social Engineering stellt ein Sicherheitsrisiko (für Privatpersonen, Unternehmen, aber auch für die nationale Sicherheit) dar. Dadurch lassen sich gezielte Phishing-Angriffe («Spear-Phishing») durchführen, um Informationen und Daten für kriminelle Absichten abzugreifen. Dadurch können Bedrohungen für die nationale Sicherheit entstehen oder mittels Verwendung der abgegriffenen Informationen anderweitige (auch Deepfake-basierte) Manipulationen politischer Entscheidungsprozesse, aber auch Erpressungskampagnen durchgeführt werden (Kötke 2021). Die Überwindung von Sicherheitsmassnahmen kann unterschiedlichen politischen oder wirtschaftlichen Zielen dienen, etwa der Sabotage kritischer Infrastruktur. In den publik gewordenen russischen Schulungsunterlagen aus den Vulkan Files wird etwa das mittlerweile stillgelegte Schweizer Atomkraftwerk Mühleberg als beispielhaftes Angriffsziel aufgeführt (Baumann u.a. 2023). Die Überwindung biometrischer Systeme mittels Deepfakes kann ausserdem der Passfälschung dienen, wodurch Personen unrechtmässigen Zugang in Staaten erhalten könnten (Europol 2022). Ein reales Beispiel für den Versuch, mittels Deepfakes das Gegenüber zu täuschen, fand im Kontext des russischen Angriffskrieges gegen die Ukraine statt. Als die Ukraine zu Beginn des Krieges auf Drohnenlieferungen

des türkischen Drohnenherstellers Bayraktar angewiesen war, gaben sich russische Agenten per Deepfake-Video als der ukrainische Ministerpräsident Denis Schmihal aus, offenbar um Waffendeals zu sabotieren. Der ukrainische Geheimdienst konnte den Deepfake-Videoanruf jedoch abfangen und demaskieren (Fulterer 2022). In einem anderen Fall, der sich später als Cheapfake herausstellte, gelang es zwei russischen Satirikern, Franziska Giffey, die Regierende Bürgermeisterin von Berlin, sowie ihr Team für eine gewisse Zeit mit einem Fake-Videoanruf des Kiewer Bürgermeisters Vitali Klitschko zu täuschen. Es stellte sich zwar später heraus, dass das Video kein Deepfake war, da der falsche Klitschko nicht KI-generiert, sondern offenbar verschiedene Videoschnipsel Klitschkos (aus altem Videomaterial) mittels Tasteneingaben gesteuert wurden. Ein solcher Angriff könnte in Zukunft allerdings auch mittels Deepfakes durchgeführt werden (Schurter 2022).

### Potenziale von Deepfakes im politischen Bereich

Trotz der vielen Gefahren und Herausforderungen haben Deepfakes das Potenzial, im politischen Bereich auch für erwünschte Zwecke verwendet zu werden. So könnten Deepfakes im Kontext von Wahlen deliberationsfördernd und für die kollektive Willens- sowie Meinungsbildung eingesetzt werden, indem sie mittels Humor und Unterhaltung den Fokus auf gesellschaftspolitische Gegebenheiten und Herausforderungen richten (Borucki u.a. 2020). So könnte eine Beschäftigung mit Deepfakes in medienpädagogischen Kontexten hilfreich sein, um Hintergründe von politischen Themen zu analysieren und darüber zu diskutieren, welche Aussagen durch Deepfakes vermittelt werden können (Appel/Prietzel 2022).[283] Dadurch könnte nicht nur erfahrbar werden, was hinter dem Phänomen Deepfakes und der Technologie steckt, sondern auch, welche Rolle manipulierte Inhalte im politischen Kontext und für die Bildung der öffentlichen Meinung spielen könnten.

Politische Akteure könnten ausserdem auf klar gekennzeichnete, satirische und humoristische Deepfakes zurückgreifen, um diese als reichweitenstarkes Mittel der Wähleransprache zu nutzen. Denn Humor und Witz gelten als wirkungsvolle Katalysatoren für öffentliche Aufmerksamkeit (Dörner/Porzelt 2016). Unabhängig vom Bildungsniveau und dem politischen Interesse der Rezipientinnen und Rezipienten könnten sich so nicht nur Chancen für die Sympathiegenerierung für die Politiker selbst eröffnen, sondern auch für wichtige politische Themen

[283] Vgl. Politik und Humor im Unterricht https://journal.hoelzel.at/wissenplus-politik-und-humor/.

(Porzelt 2015: 5). Dennoch besteht die Gefahr, dass satirische Deepfakes zu Täuschungszwecken eingesetzt werden und aufseiten politischer Handlungsträger, aber auch der Zivilgesellschaft Schaden anrichten, der sich über die Wahlbeeinflussung bis hin zur Erosion des Vertrauens in politische Institutionen und den Journalismus erstrecken kann. Dies bildet auch den Kontext der bisherigen Nutzung von Deepfakes in der Schweiz, etwa im Sommer 2023 durch die FDP in Form der Verwendung eines KI-generierten Bildes, das Klimaaktivisten bei der Blockierung eines Rettungsfahrzeugs darstellt. Die Verwendung des Deepfake-Bildes zog Kritik auf sich, weil es die Klimaaktivisten auf eine Weise darstelle, die nicht der Realität entspreche. Bislang sind nämlich keine realen Fälle aus der Schweiz bekannt, in denen ein Rettungsfahrzeug seitens Klimaaktivisten blockiert wurde (Helfenberger 2023). Ende September 2023 verpflichtete sich eine Allianz aus Mitte- und Linksparteien (Grünen, SP, Mitte, EVP und GLP) freiwillig dazu, die Verwendung von KI im Schweizer Parlamentswahlkampf 2023 zu deklarieren und keine KI-Inhalte mit Bild oder Ton für Negativkampagnen zu missbrauchen. Die FDP und SVP enthielten sich mit der Begründung, dass die Grenzziehung zwischen erlaubter Wahlwerbung und Satire sowie Negativkampagnen schwierig sei (Strasser 2023). Weniger später verbreitete SVP-Nationalrat Andreas Glarner im Oktober 2023 ein satirisches Deepfake-Video, das die Grünen-Nationalrätin Sibel Arslan zeigt, wie sie die Ausschaffung «aller kriminellen Türken» fordert und zur Wahl von Glarner aufruft (Niessner u.a. 2023). Nachdem Arslan Glarner zivilrechtlich wegen der Verletzung ihrer Persönlichkeitsrechte verklagt hatte, verurteilte das zuständige Gericht Glarner Ende 2023 zur Zahlung einer Geldstrafe. Wie eine mehrheitlich akzeptierte politische Nutzung von Deepfakes aussehen kann, wird sich also noch zeigen müssen.

### 6.1.3. Zwischenfazit: Einsatz von Deepfakes in der Schweizer Politik

In vorangegangenen Kapiteln wurden die mit der Digitalisierung einhergehenden allgemeinen Herausforderungen aufgezeigt, die sich innerhalb der politischen Öffentlichkeit für Gesellschaft, Individuum, politische Akteure und Parteien ergeben. Hierfür wurden zunächst die Implikationen von Deepfakes für die Politik anhand der aktuellen Literatur zu Deepfakes und Desinformation aufgearbeitet.

Insgesamt zeigt sich, dass zahlreiche mögliche Auswirkungen von Deepfakes auf die Politik bereits Gegenstand von verschiedenen wissenschaftlichen Publikationen waren. Allerdings diskutieren unterschiedliche Texte unterschiedliche Gefährdungsaspekte in je verschiedener Weise. Zudem besteht ein Fokus insb. auf Aspekten der Gefährdung der Demokratie insgesamt und der Manipulation

von demokratischen Wahlen mittels Deepfakes. Insofern fehlt ein Gesamtüberblick der durch Deepfakes hervorgerufenen Herausforderungen für die Politik. Sofern Szenarien diskutiert werden, beruhen diese nicht auf einer stringenten Struktur. Arbeiten mit Bezug zur Schweizer Politik fehlen hierbei gänzlich.

Das vorliegende Kapitel hat daher zum Ziel, diese Forschungslücke zu adressieren. Dies erfolgt mittels Bearbeitung der folgenden Forschungsfragen:

**FF 5.1:** Welche Rolle spielen Deepfakes gegenwärtig und künftig in der Schweizer Politik?

**FF 5.2:** Wie können die von Deepfakes in der Politik ausgehenden Herausforderungen adressiert werden?

**FF 5.3:** Welche Handlungsoptionen bieten sich der Politik und anderen Akteuren?

Zu diesem Zweck wurde im nächsten Schritt zunächst eine Umfrage unter Schweizer Parlamentarierinnen und Parlamentariern sowie Institutionen der Bundesverwaltung durchgeführt (vgl. Kapitel 6.2). Abschliessend wurden die Ergebnisse aus der Diskussion des Forschungsstands und der Umfrage in Form von Szenarien zu Deepfakes in der Politik zusammengeführt (vgl. Kapitel 6.3).

## 6.2. Umfrage im Schweizer Parlament und der Bundesverwaltung

### 6.2.1. Methodisches Vorgehen

*Zielsetzung und Methodenwahl*

In den vorhergegangenen Kapiteln wurde eine Eingrenzung und Strukturierung des Status quo von Deepfakes in der Schweizer Politik herausgearbeitet. Diese explorative und damit qualitativ angelegte Phase legte den Schwerpunkt auf die Makrosicht über den Forschungsstand zu Deepfakes in der Politik und bildet die Basis für eine empirische Untersuchung. Diese erfolgte mittels einer Onlinebefragung Schweizer Parlamentarierinnen und Parlamentarier und thematisch einschlägigen Institutionen der Bundesverwaltung. Das quantitative Vorgehen ermöglichte eine inhaltlich breite Erhebung, wobei der Status quo der Rolle von Deepfakes in der Politik quer durch das Parteienspektrum des Schweizer Parlaments erhoben wurde.

*Zielgruppe und Fallauswahl*

Die Zielgruppe der Onlinebefragung setzte sich zum einen aus den Mitgliedern von Nationalrat und dem Ständerat (Akteure Politik) und zum anderen aus thematisch einschlägigen gesellschaftspolitischen Initiativen aus dem Bereich der Digitalisierung sowie Bundesministerien (Akteure Verwaltung) zusammen. Die E-Mail-Kontaktadressen der politischen Amtsträger sind auf der offiziellen Internetseite der Bundesversammlung zugänglich und wurden von dort bezogen. Das Kriterium für die Auswahl der Befragten für den Bereich der Verwaltung war die thematische Ausrichtung. Es wurden Ministerien kontaktiert, die aufgrund der thematischen Schwerpunktsetzung auf Themen wie Digitalisierung und Medienregulierung einschlägig erschienen. Darüber hinaus wurden gesellschaftspolitische Initiativen der Schweiz mit dem Fokus auf Digitalisierungsthemen kontaktiert. Die Zusammensetzung des Samples und die Kontaktaufnahme wurden eng mit TA-SWISS abgestimmt. Dem Anschreiben konnte dank der Unterstützung seitens der TA-SWISS-Projektleitung ein Unterstützungsschreiben von Walter Thurnherr, Bundeskanzler der Schweiz, beigelegt werden.

**Tabelle 5:** Zielgruppe der Onlinebefragung zum Thema Deepfakes in der Politik (Stand: März 2023 gemäss parlament.ch)

| Organ/Institution | Anzahl der kontaktierten Personen |
|---|---|
| Nationalrat | 199 |
| Ständerat | 45 |
| Politische Initiativen | 6 |
| Bundesverwaltung | 26 |
| Gesamt (n) | 276 |

Anders als bei den kontaktierten Personen aus dem Schweizer Parlament wurde bei den Akteuren aus der Verwaltung und den politischen Initiativen kein personalisierter Link zur Befragung versendet. Nicht aus Gründen der Anonymität der Befragungssituation, diese war zu jedem Zeitpunkt gegeben, sondern um Gebrauch von einem möglichen Schneeballeffekt und aktiver Rekrutierung der angeschriebenen Erstkontakte machen zu können (Möhring/Schlütz 2019: 144 f.). Die Akteure aus den beiden genannten Gruppen wurden im Anschreiben der Befragung explizit um die Weiterleitung der Befragung an interessierte Dritte gebeten. Insgesamt ergab sich für die vom 02. Februar bis 20. Februar 2023 laufende Onlineumfrage ein Gesamtsample von n = 276 (vgl. Tabelle 5). Um die

Rücklaufquote zu erhöhen, wurden in der Laufzeit der Befragung in zeitlichen Abständen von je einer Woche Erinnerungsmails an das gesamte Sample gesendet.

*Leitfaden und Programmierung*

Als Einstieg zum Onlinefragebogen wurde ein kurzer thematischer Überblick vor die Umfrage gesetzt. Der Leitfaden für die Befragung wurde auf Basis des Forschungsstandes und des forschungsleitenden Erkenntnisinteresses konzipiert. Die Onlineumfrage gliederte sich in die Bereiche:

- Einführung
- Relevanz von Deepfakes für den jeweiligen Arbeitsbereich
- Chancen und Risiken von Deepfakes
- Schutzmassnahmen gegen Deepfake-Angriffe

Ziel der Umfrage war es, ein Meinungsbild zum Thema Deepfakes bei Akteuren aus Politik und Verwaltung zu erhalten. Hierfür wurde mit verschiedenen Fragetypen gearbeitet, die von Matrixantworten, über individuelle Einschätzungen zur Wahrnehmung (Likert-Skala) oder Freitextantworten reichten, um ein möglichst umfangreiches Bild an Meinungen, Kommentaren und Hinweisen zu ermöglichen (siehe Onlineappendix[284]). Die technische Umsetzung der Onlineumfrage erfolgte anhand des Befragungstools *EFS Survey*. Die Fragebögen waren in deutscher und französischer Sprache programmiert. Darüber hinaus wurde den Teilnehmenden angeboten, sie über die Studienveröffentlichung und Ergebnisse zu informieren.

*Auswertung und Reflexion*

Die Gütekriterien quantitativer Sozialforschung *Objektivität, Reliabilität und Validität* waren handlungsleitend für den gesamten Prozess der Entwicklung des methodischen Designs, der Datenerhebung und -auswertung (Krebs/Menold 2019). Die Auswertungsmethode wurde nach genauer Betrachtung der Fragen und Antwortmöglichkeiten ausgewählt. Bei Befragungen unter Parlamentariern stellt sich häufig das Problem einer geringen Rücklaufquote (Debus/Bäck 2014), so auch im vorliegenden Fall: n = 29 (Politik: 23 Teilnehmende; Verwaltung: 6 Teilnehmende). Angesichts der geringen Rücklaufquote war es schwierig, einen umfassenden Überblick über die Rolle und den Status quo von Deepfakes

[284] https://zenodo.org/records/10848048.

in der Politik zu erhalten. Die niedrige Beteiligung gilt es kritisch bei der Interpretation der Ergebnisse zu reflektieren (Non-Response-Bias). Die Ursachen für die geringe Rücklaufquote sind nicht valide erklärbar, da keine Befragung diesbezüglich durchgeführt wurde. Die Antworten der Teilnehmenden lassen jedoch darauf schliessen, dass dem Thema Deepfakes in der Schweizer Politik noch nicht ausreichend viel Raum geboten wird, was an mangelndem Interesse bzw. unzureichendem Wissen über die Implikationen von Deepfakes liegen kann.[285]

Aufgrund der geringen Zahl an Antworten aus der Verwaltung verwenden wir lediglich die eingegangenen Antworten auf Freifelder und beschreiben Häufigkeiten. Aussagen im Sinne einer Inferenzstatistik können aufgrund der Datenbasis nicht getätigt werden.

Die Befragung kann jedenfalls kein statistisch fundiertes Abbild zum Status quo von Deepfakes in der Schweizer Politik leisten. Die Ergebnisse ermöglichen jedoch einen ersten Einblick in Tendenzen der Relevanz, aber auch blinde Flecken des Themas in der Schweizer Politik. Die Befragungsergebnisse können somit bei der Relevanzbestimmung der Szenarien im folgenden Kapitel dienlich sein und eine Diskussionsgrundlage in Bezug auf Schutzmassnahmen und Handlungsempfehlungen für den Umgang mit Deepfakes liefern.

**Tabelle 6:** Fraktionszugehörigkeit der Studienteilnehmenden

| Fraktionszugehörigkeit der Teilnehmenden | Anzahl Teilnehmende | Anzahl Sitze in Bundesversammlung |
|---|---|---|
| Grüne Fraktion | 8 | 35 |
| Die Mitte-Fraktion | 3 | 44 |
| Grünliberale Fraktion | 3 | 16 |
| Sozialdemokratische Fraktion | 6 | 48 |
| FDP-Liberale Fraktion | 3 | 41 |
| Fraktion der Schweizerischen Volkspartei | 0 | 62 |

[285] Weitere mögliche Erklärungen sind: zunächst die steigende Arbeitslast des Parlaments (gemessen an Vorstössen), auch im Hinblick auf die allgemeine Flut an Anfragen (Bühler (2022), die hiermit verbundenen geringen zeitlichen Ressourcen von politischen Handlungsträgerinnen und Handlungsträgern und darüber hinaus die Komplexität und Sensibilität des Themas Deepfakes mit Blick auf die wenige Zeit, die der Zielgruppe zum Einstieg in das Thema und der Beantwortung der Fragen bleibt.

**Tabelle 7:** Kammerzugehörigkeit der Studienteilnehmenden

| Kammerzugehörigkeit der Teilnehmenden | Anzahl |
|---|---|
| Nationalrat (199 Mitglieder) | 17 |
| Ständerat (45 Mitglieder) | 6 |

### 6.2.2. Resultate

Die allgemeine Rolle von Deepfakes in der Schweizer Politik scheint mit Blick auf die Umfrageergebnisse noch keine erhebliche zu sein (vgl. Abbildung 24). Hierbei gaben die Befragten an, dass Deepfakes ihnen vor allem aus Medienberichten bekannt sind, in denen über Fälle von Deepfakes berichtet wurde, die auf internationaler Ebene für Aufmerksamkeit gesorgt haben. Konkrete eigene Erfahrungen (insb. als Opfer) wurden nicht genannt. Doch auch diejenigen Befragten, die angegeben hatten, dass Deepfakes noch kein Thema in der Schweizer Politik seien (n = 11), äusserten mit grosser Mehrheit den Wunsch, dass dem Thema mehr politische Aufmerksamkeit zuteilwerden sollte (vgl. Abbildung 25).


**Abbildung 24:** Rolle von Deepfakes in der Schweizer Politik (n = 23)

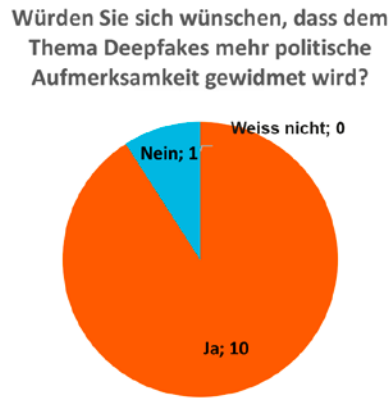


**Abbildung 25:** Wird mehr politische Aufmerksamkeit für das Thema Deepfakes gewünscht? (n = 11)

*Relevanz von Deepfakes für den Arbeitsbereich der Befragten*

Die Mehrheit der Befragten (n = 14) gab an, dass Deepfakes auf ihrer Arbeitsebene kein Thema sind. Fünf Personen äusserten sich dahin gehend, dass

das Thema Deepfakes in ihrem jeweiligen Arbeits- und Verantwortungsbereich durchaus wichtig ist und sie in ihrer politischen Arbeit beschäftigt. Unter anderem stelle sich die Frage nach der redaktionellen Verantwortung von Plattformen und deren Regulierung, juristischer Beweisregeln und den Auswirkungen von Deepfakes für diplomatische Beziehungen. Hierbei wird insbesondere Vertrauen im Zusammenhang mit durch die Technologie verbreiteten Manipulationen thematisiert: Nach Aussage eines Teilnehmers werde das für den effektiven politischen Betrieb erforderliche Einholen von Informationen seitens der Parlamentarier erschwert, wenn die Parlamentarier sich nicht mehr sicher sein können, dass die eingehenden Informationen auch vertrauenswürdig sind. Als mögliche Herausforderung wurde auch die Torpedierung des eigenen Wahlkampfes infolge irreführender Deepfakes benannt. Ein Teilnehmer gab an, dass ein frühzeitiges Monitoring der Trends zu Deepfakes sowie der konkreten Risiken von Videoidentifikationsverfahren wichtig seien. Der Austausch über die Relevanz von Deepfakes in der Politik finde zwischen den politischen Handlungsträgern mitunter in institutionalisierten Arbeitsgruppen statt.

Hierbei stellt sich eine Person die Frage, ob das gegenwärtige Strafrecht ausreichend ist und wie den Herausforderungen in der politischen Öffentlichkeit wie Desinformation, Identitätsdiebstahl und der Verbreitung von «erniedrigenden Fakes» begegnet werden kann. Drei Befragte äusserten auch Sorgen, dass «man selbst nicht in die Falle geraten möchte», eine Korrektur von falschen Informationen aufwendig wäre und dennoch ein etwaiger Reputationsschaden entstehen könne.

*Risiken und (kaum) Chancen von Deepfakes*

In der Einschätzung der konkreten Risiken (vgl. Abbildung 26) wurden als grösste Risiken die Wirkungen von Deepfakes für die Schweizer Demokratie sowie die Risiken für die politischen Institutionen, welche die Befragten vertreten, genannt (je 20 Nennungen). 18 Befragte gaben an, dass es ein ernst zu nehmendes Risiko darstelle, wenn Deepfakes über sie selbst kursierten. 16 Teilnehmende erkennen auch Risiken für die internationalen Beziehungen zu anderen Staaten und 15 Teilnehmende sahen eine Gefahr, dass auch sie selbst durch Deepfakes getäuscht werden könnten.


**Abbildung 26:** Wahrgenommene Risiken von Deepfakes in der Schweizer Politik (Mehrfachnennung möglich)

Bei der Frage, wie hoch bzw. niedrig die Eintrittswahrscheinlichkeit der jeweiligen Risiken eingeschätzt wird, zeigte sich ein eindeutiges Bild (vgl. Abbildung 27): Die Eintrittswahrscheinlichkeit der Risiken für die Schweizer Demokratie wurde mehrheitlich als am höchsten eingestuft. Je sechs Teilnehmende sahen eine sehr hohe bzw. hohe Wahrscheinlichkeit, dass ein solches Risiko tatsächlich eintritt. Auch die Eintrittswahrscheinlichkeit der Risiken für die Institutionen, die durch die Befragten vertreten werden, wurde von zwei Befragten als sehr hoch und von acht Befragten als hoch bewertet, während lediglich eine Person die Wahrscheinlichkeit als niedrig einschätzte. Ein anderes Bild ergibt sich bei der Bewertung der Eintrittswahrscheinlichkeit des Risikos, selbst mittels eines Deepfakes getäuscht zu werden, dass ein Deepfake über einen selbst kursiert sowie der Risiken für internationale Beziehungen. Eine Mehrheit der Befragten war hier jeweils der Ansicht, dass die Wahrscheinlichkeit als niedrig einzustufen ist.

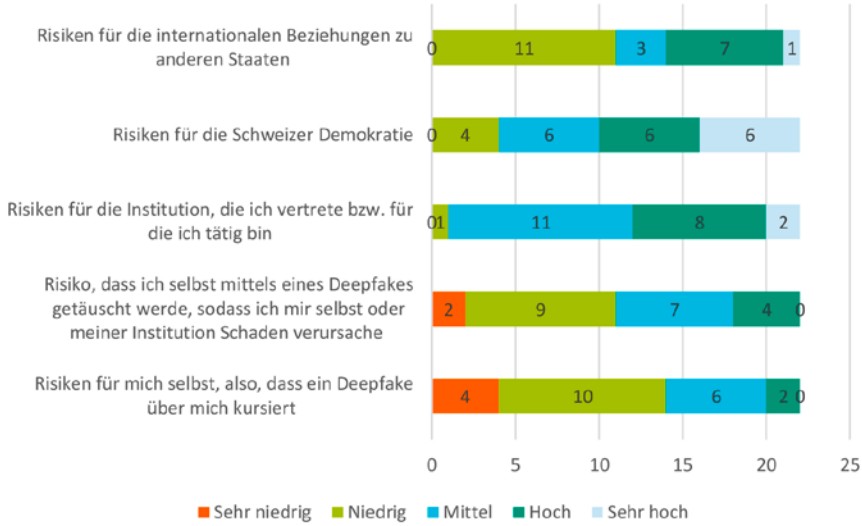


**Abbildung 27:** Einschätzung der Eintrittswahrscheinlichkeit von Deepfake-Risiken

Ein gemischtes Bild zeigte sich bei den Antworten auf die Frage, inwiefern die benannten Risiken im Arbeitsbereich der Befragten bereits konkret diskutiert werden (vgl. Abbildung 28). Elf Teilnehmende gaben an, dass keine Diskussionen über die Risiken stattfänden. Eine Person gab an, dass alle relevanten Risiken diskutiert würden, und neun weitere, dass die Risiken teilweise Diskussionsgegenstand seien.

So wurde etwa auf Vorstösse unter dem Oberthema «Desinformation» verwiesen (Das Schweizer Parlament 2021). Auch ein Bericht zur Plattformregulierung wurde als wichtiger Diskussionsanstoss benannt. Etwa, dass der Bundesrat das Departement für Umwelt, Verkehr, Energie und Kommunikation (UVEK) beauftragt hatte, bis Ende März 2024 eine Vernehmlassungsvorlage zur Regulierung der Kommunikationsplattformen auszuarbeiten (Bundesrat 2023a). Die Bedeutung und Finanzierung des öffentlich-rechtlichen Rundfunks und das Verhältnis zu sozialen Medien ist in diesem Zusammenhang ein Aspekt, der von einem politischen Handlungsträger schon heute im Arbeitsbereich diskutiert wird. Ebenfalls diskutiert wird in den Arbeitsbereichen der teilnehmenden Politikerinnen und Politikern das Szenario, selbst Opfer von Deepfakes zu werden (Identitätsdiebstahl), indem Aussagen sowie biometrische Merkmale verändert werden. Dass Deepfakes auch mit Chancen verbunden werden, zeigt ein Teilnehmer, der auf die Potenziale der Sprachübersetzung via Deepfakes hinweist. Schulungsvideos, Vorschläge für Textdokumente und eine Real-time-Übersetzung von Videokonferenzen werden ebenfalls als Beispiele eingebracht.

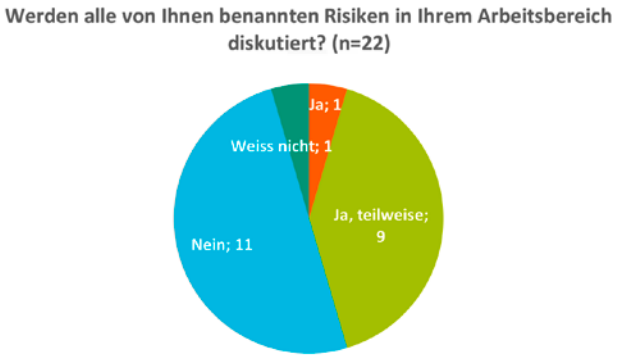


**Abbildung 28:** Diskussion benannter Risiken im jeweiligen Arbeitsbereich

Eindeutig war auch, dass von keinem Teilnehmenden Chancen in Bezug auf Deepfakes in der Politik erkannt wurden. Während vier Personen mit «weiss nicht» antworteten, wurden mögliche Chancen von der Mehrheit der Befragten (19) ausgeschlossen.

*Schutzmassnahmen vor Deepfake-Angriffen*

Die Mehrheit der Befragten (14 Teilnehmende) gab an, dass sie bislang keine konkreten Schutzmassnahmen ergriffen haben, um die von Deepfakes ausgehenden Risiken einzudämmen, während lediglich zwei Teilnehmende die Frage bejahten (vgl. Abbildung 29). Gefragt nach der Art der getroffenen Massnahmen, gaben beide Personen individuelle Schutzmassnahmen durch sie selbst an. Ausserdem wurde auf die Aufklärung der Bevölkerung, Schutzmassnahmen durch eigene Teammitarbeitende und durch die eigene Institution bzw. Partei verwiesen.


**Abbildung 29:** Schutzmassnahmen gegen Deepfakes in der Schweizer Politik

### 6.2.3. Zusammenfassung

Insgesamt gab eine Mehrheit der Befragten an, dass Deepfakes entweder bereits ein Diskussionsthema im politischen Betrieb sind, oder sie äusserten den Wunsch, dass dem Thema mehr Aufmerksamkeit gewidmet werden sollte. Es sollte jedoch bedacht werden, dass an der Umfrage womöglich überwiegend jene Personen teilgenommen haben, die dem Thema ohnehin eine hohe Bedeutung beimessen.

Hinsichtlich der offenen Frage, inwiefern Deepfakes die eigene politische Arbeit berühren, zeigte sich, dass zahlreiche der in der Fachliteratur erwähnten Aspekte von Deepfakes auch in der (politischen) Praxis bereits diskutiert werden oder auf der Agenda der Befragten stehen. Darunter fallen Themen wie Plattformregulierung und -verantwortung, die Verlässlichkeit juristischer Beweisregeln, Herausforderungen bezüglich des geltenden Strafrechts, die Vertrauenswürdigkeit eingehender Informationen, die Risiken für Wahlkämpfe, für internationale Beziehungen und dass man durchaus auch selbst mittels eines Deepfakes in die Irre geführt werden könnte.

Bei den wahrgenommenen Chancen und Risiken zeigte sich ein eindeutiges Bild: Die Befragten sahen fast ausschliesslich Risiken. Bedenken fokussierten sich insbesondere auf die Gefahren für die Schweizer Demokratie und die politischen Institutionen, die von den Teilnehmenden vertreten werden. Hinsichtlich beider Aspekte sahen die Befragten zudem mehrheitlich eine hohe oder sehr hohe Eintrittswahrscheinlichkeit. Das Risiko, dass ein Deepfake über einen selbst kursiert bzw. dass man mittels eines Deepfakes getäuscht wird sowie die aus Deepfakes resultierenden Risiken für internationale Beziehungen wurden ebenfalls häufig als relevantes Risiko benannt. Die Mehrheit der Teilnehmenden war allerdings zugleich der Ansicht, dass die Wahrscheinlichkeit des Eintritts derartiger Risiken als eher niedrig bis sehr niedrig einzustufen ist.

Die Befragung zeigt schliesslich auch eine dahin gehende Tendenz, dass gegenwärtig noch zu selten konkrete Schutzmassnahmen gegen Deepfakes ergriffen werden.

## 6.3. Szenarien zu Deepfakes in der Politik

Das vorliegende Kapitel entwirft Szenarien zu Deepfakes in der Politik, um die von Deepfakes ausgehenden Gefahren für die Schweizer Politik zu konkreti-

sieren. Im Folgenden werden die gewonnenen Erkenntnisse aus den vorangegangenen Kapiteln zusammengeführt und in Form einer zusammenfassenden Gesamtübersicht unterschiedlicher Gefährdungsszenarien dargestellt. Die ausführlichen Szenarien sind im Anhang zu finden (vgl. A.5).

Die Szenarien fussen auf der Zusammenführung der relevanten Forschungsliteratur (Verhulst 2020; van Huijstee u.a. 2021; Bateman 2020; Bundesministerium für Inneres 2022; Kalpokas/Kalpokiene 2022; Collins 2019; Chesney/Citron 2018; Westerlund 2019) mit den Erkenntnissen aus der Befragung. Bei der Diskussion der Szenarien wird für einen besseren Überblick zwischen den Ebenen Individuum, Organisation und Gesellschaft unterschieden – auch wenn zu erwarten ist, dass bei vielen Szenarien ebenenübergreifende Wirkungen eintreten können. Rufschädigung, beispielsweise, kann Ergebnis unterschiedlicher Einsatzweisen von Deepfakes sein: Ein Angreifer kann bewusst die Schädigung des Rufs eines bestimmten Politikers beabsichtigen, um z.B. dessen (Wieder-)Wahl zu sabotieren. Wenn dessen Partei oder das bekleidete staatliche Amt bzw. die staatliche Institution, die der Politiker in diesem Fall vertritt, Schaden nehmen, ist dies als Kollateralschaden des auf den Politiker abzielenden Deepfake-Angriffs zu betrachten. Die Beschädigung des Rufs einer Politikerin oder eines Politikers kann aber auch Vehikel sein, um dessen Partei zu schaden und dadurch beispielsweise einen Wahlausgang in der Breite zu beeinflussen oder sogar um das Vertrauen in das politische System als Ganzes zu untergraben. Zugleich können auch zu Unterhaltungszwecken erstellte Deepfakes seitens der Bevölkerung zu einer Rufschädigung führen – also auch wenn eine bewusste Intention zur Schädigung nicht vorliegt.

Für den Zweck des vorliegenden Kapitels, die gegenwärtige und künftige Rolle von Deepfakes für die Schweizer Politik zu untersuchen, orientiert sich die Bestimmung der jeweiligen Szenarien an dem übergeordneten Effekt eines Deepfakes. Wenn also der übergeordnete Effekt die Untergrabung des Vertrauens in das politische System ist, dann sind die Beschädigung des Rufs eines Politikers und von dessen Partei als untergeordnete Effekte zu betrachten (Collins 2019: 11).

### Angriffstypen und erforderlicher Ressourcenaufwand

Im Folgenden unterscheiden wir fünf grundlegende Deepfake-basierte Angriffstypen. Jedem dieser Angriffstypen lassen sich unterschiedliche Techniken (vgl. Kapitel 2) zuordnen. Die Nummerierung orientiert sich an der kurz-, mittel- bzw. langfristigen Realisierbarkeit und des für den jeweiligen Angriffstyp erforder-

lichen Ressourcenaufwands, das heisst, gefälschte Aussagen bilden den Angriffstyp, der schon heute möglich und in rudimentärer Qualität bereits mit einem geringen Aufwand realisierbar ist. Vollständig synthetische Inhalte stellen dagegen einen Angriffstyp dar, der erst in einigen Jahren möglich sein wird. In jedem Angriffsszenario können ein oder mehrere Angriffstypen kombiniert zur Anwendung kommen.

1. **Gefälschte Aussagen oder Handlungen**: Der bekannteste und inzwischen bereits mehrfach eingesetzte Deepfake-Angriffstyp ist die Fälschung von Aussagen mittels eines Deepfakes. Durch Manipulation des Gesichtsausdrucks und Gesichtsaustausch lassen sich authentisch wirkende Fälschungen von Personen generieren. Kombiniert mit der Technik des Stimmenklonens können der gefälschten Person schliesslich irreführende Worte in den Mund gelegt werden. Mittels moderner Chatbots, wie ChatGPT, ist es inzwischen auch möglich, diese Worte auf automatisierte Weise so klingen zu lassen, dass sie der Wortwahl des Originals möglichst nahekommen. Mit der zunehmenden Qualität von Deepfakes wird es in Zukunft sogar möglich sein, nicht nur den Gesichtsausdruck und das Gesprochene, sondern mittels der Technik des Ganzkörperpuppenspiels auch die Fälschung von körperlichen Handlungen durchzuführen.

2. **Social-Engineering-Angriff**: Der zweite Angriffstyp ist der Social-Engineering-Angriff mittels Deepfakes. Beim Social Engineering nutzen Angreifer menschliche Eigenschaften wie Hilfsbereitschaft, Vertrauen, Angst oder Respekt vor Autoritäten für die Manipulation der Opfer. So kann ein Opfer dazu verleitet werden, vertrauliche Informationen preiszugeben, Sicherheitsfunktionen zu deaktivieren, Überweisungen zu tätigen oder Schadsoftware auf einem privaten oder beruflichen Gerät zu installieren (Wang u.a. 2021). Die bekannteste und bereits mehrfach eingesetzte Form eines solchen Social-Engineering-Angriffs ist das Deepfake-Voice-Phishing, bei der eine geklonte Stimme zur Überlistung des Gegenübers genutzt wird. Künftig sind aber auch auf audiovisuellen Deepfakes basierende Social-Engineering-Angriffe wahrscheinlich, insb. mittels der Technik der Manipulation des Gesichtsausdrucks, aber auch des Gesichtsaustauschs.

3. **Überwindung von Sicherheitsmassnahmen**: Prinzipiell ist die Überwindung von Sicherheitssystemen mittels Deepfakes heute schon möglich. Mittels einer geklonten Stimme lassen sich etwa Authentifizierungssysteme überwinden, die mit Stimmerkennung arbeiten. Mittels Gesichtsmorphing und Gesichtsaustausch besteht diese Möglichkeit auch bei gesichtsbiometrischen Authentifizierungssystemen.

4. **Synthetische Social Botnets**: (Social) Botnets und Fake-Accounts stellen bereits seit vielen Jahren ein grosses Problem für die Betreiber grosser Onlineplattformen dar. Dieser Angriffstyp wird in Richtung sog. synthetischer Social Botnets weiterentwickelt. Manipulierte oder synthetisch generierte Audio-, Bild- und Video-Deepfakes und KI-generierte Texte können dazu genutzt werden, Social-Media-Profile von nicht existierenden Menschen zu erstellen und zu betreiben. Mittels Gesichtsgeneratoren können beispielsweise synthetische Gesichter erzeugt werden, die schwieriger zu erkennen sind als herkömmliche Bots, die Fotos real existierender Personen nutzen. Mithilfe von modernen KI-Methoden könnten Social Bots untereinander aber auch mit echten Social-Media-Nutzenden so natürlich kommunizieren, dass sie kaum oder gar nicht als Bots erkannt werden können. Dabei können die Social Bots jeweils eine unterschiedliche «Persönlichkeit» besitzen und authentisch in ihrer Rolle agieren. Dies würde sowohl die Glaubwürdigkeit der Bots erhöhen und damit ihre Detektion erschweren. Auf diese Weise könnten die Bots auch in ganz unterschiedliche Bevölkerungsgruppen hineinwirken, sodass die Wirkung eines Deepfake-Angriffs durch höchstmögliche Streuung maximiert wird. Ein *bildungsbürgerlicher Bot*, ein *links- oder rechtsgerichteter Bot* würden dann jeweils unterschiedliche Bevölkerungsgruppen adressieren. Sobald der Aktivierungsbefehl kommt, könnten sie dann auf koordinierte Weise Postings mit dem Deepfake absetzen, während sie einerseits in ihrer zuvor definierten Rolle verbleiben und andererseits klassische Botnet-Angriffsdetektoren umgehen, weil nicht alle Postings zeitgleich abgesetzt würden, sondern mit ausreichend viel zeitlichem Abstand (van Huijstee u.a. 2021: 51; Bateman 2020: 7).

5. **Deepfakes von Objekten und Situationen/vollständig synthetische Inhalte**: In Zukunft wird es schliesslich auch möglich sein, audiovisuelle Deepfakes von Objekten und Situationen zu erstellen, die so nie existiert haben. Schon heute wird an der Übertragung der von KI-basierten Bildgeneratoren bekannten Möglichkeiten auf Videogeneratoren geforscht (Meta 2022). Sobald diese Technologie verfügbar ist, werden Angreifer in der Lage sein, beliebige audiovisuelle Inhalte mittels Eingabe von Textbefehlen zu generieren. Die positiven wie negativen Möglichkeiten dieser Technik scheinen grenzenlos: gefälschte Naturkatastrophen, gefälschte öffentliche Demonstrationen, gewalttätige polizeiliche Übergriffe, Terroranschläge oder zwischenstaatliche Überfälle – um nur einige Einsatzmöglichkeiten zu nennen.[286]

[286] Zu erwarten ist, dass klassische Deepfake-Techniken, wie die Manipulation des Gesichtsausdrucks, das Gesichtsmorphing usw., sukzessive durch KI-basierte Videogeneratoren ersetzt werden.

**Tabelle 8:** Zuordnung von Deepfake-Techniken (y-Achse) zu Deepfake-Angriffstypen (x-Achse)

| | **Fälschung von Aussagen und Handlungen** | **Social-Engineering-Angriff** | **Überwindung von Sicherheitsmassnahmen** | **Synthetische Social Botnets** | **Deepfakes von Objekten und Situationen** |
|---|---|---|---|---|---|
| Manipulation des Gesichtsausdrucks | x | x | | | |
| Gesichtsmorphing | | | x | (x) | |
| Gesichtsaustausch | x | x | | | |
| Gesichtsgenerierung | | | | x | |
| Ganzkörperpuppenspiel | x | | | | |
| Klonen der Stimme | x | x | | | |
| KI-generierter Text (Chatbots, usw.) | x | | | x | |
| KI-basierte Videogeneratoren | x | x | X | | x |

### Adressatenkreis und Verbreitungsweise

Eine weitere wichtige Unterscheidung zwischen den Deepfake-Gefährdungsszenarien liegt im Adressatenkreis, dem das Deepfake dargestellt wird, sowie der Verbreitungskanal eines Deepfake-basierten Inhalts. Die nähere Betrachtung des Adressatenkreises und der Verbreitungsweise ist relevant, weil vom Adressatenkreis die konkreten Kommunikationskanäle abhängen, die zur Aussendung bzw. Verbreitung eines Deepfake-Inhalts genutzt werden.

Bei Szenarien mit einem engen Adressatenkreis kann es dem Angreifer etwa darum gehen, durch einen privaten Kanal direkten Kontakt zur Zielperson aufzunehmen, um diese mittels eines Social-Engineering-Angriffs in die Irre zu führen. Im Falle der Nutzung eines Deepfakes zum Zwecke der Erpressung wird ebenfalls ein direkter Kanal zur Zielperson hergestellt, um sie unter Androhung der weiten Streuung des entsprechenden Deepfakes zu bestimmten Handlungen zu zwingen. Bei Szenarien mit einem weiten Adressatenkreis ist der Einsatz eines Deepfakes auf die möglichst breite Streuung des entsprechenden Inhalts ausgelegt. Beispielsweise würde im Falle eines Eingriffsversuchs in eine demokratische Wahl, bei dem eine Politikerin bei einer vermeintlich illegalen Handlung dargestellt wird, das entsprechende Deepfake möglichst breit gestreut werden, um die gesellschaftliche Wirkung zu erhöhen (Bateman 2020: 6).

### 6.3.1. Kurzzusammenfassung der Szenarien

Auf Basis der Literatur haben wir elf Szenarien identifiziert, wie Deepfakes im politischen Kontext Verwendung finden können. Die ausführliche Diskussion der Szenarien (vgl. im Anhang A.5) zeigt, wie Deepfakes für zahlreiche intendierte wie nicht intendierte Zwecke, die mit einem Schaden einhergehen, eingesetzt werden können (für einen Überblick, vgl. Tabelle 9).

Hierzu zählen die Erpressung bzw. Einschüchterung und die Rufschädigung von Politikerinnen und Politikern, Funktionsträgern usw. Auch können Deepfakes dazu beitragen, zum Hass aufzustacheln und zur Gewalt aufzurufen. Zudem eignen sich Deepfakes zur Rufschädigung politischer Institutionen und zur Erbeutung von vertraulichen Informationen, indem Mitarbeitende und ggf. biometrische Authentifizierungssysteme überlistet werden.

Der Grossteil der Szenarien setzt sich schliesslich mit den vieldiskutierten möglichen Effekten auf gesellschaftlicher Ebene auseinander. Neben der Beeinflussung von Wahlen und der Beeinflussung von politischen Entscheidungsprozessen werden auch Szenarien zur Verschärfung sozialer Spannungen, Beschädigung der Demokratie, Gefährdung der öffentlichen Sicherheit und Beeinflussung der internationalen Beziehungen erörtert.

Aufgrund der schwierigen Plausibilisierbarkeit machen wir keine Aussagen über Eintrittswahrscheinlichkeiten. Die Szenarien zeigen allerdings, dass sich mit wenig Aufwand produzierte Deepfakes im politischen Kontext nutzen lassen. Hierzu zählen insbesondere Deepfake-Bilder, aber zunehmend auch einfach

produzierte Deepfake-Videos. Verschiedene reale Nutzungen von Deepfakes (vgl. auch die Diskussion in Kapitel 6.1) zeigen ausserdem, dass Einsatzzwecke wie Betrug und Rufschädigung bereits heute naheliegend sind.

Klar wird in den Szenarien zudem, dass die fortgeschrittenen Deepfake-Nutzungen ein enormes Know-how und einen hohen Ressourcenaufwand erfordern. Die Sammlung persönlicher Daten (etwa Gesichtsfotos und Stimmsamples), deren Nutzung zum Training von Deepfake-Software bis zur schliesslichen Verwendung der Deepfake-Software zum Erstellen glaubwürdig anmutender Videoszenen sind bereits Arbeitsschritte, die nicht ohne Weiteres von Laien durchgeführt werden können.

Je nachdem, ob es das Ziel ist, den produzierten Inhalt einer einzelnen Person, einer kleineren Personengruppe oder möglichst vielen Menschen zu zeigen, werden entsprechend der Grösse des Adressatenkreises und der Schwierigkeit, einzelne Personen oder Gruppen zu erreichen, weitere Ressourcen erforderlich. Gerade der Einsatz von Social Bots und ausgeklügelte Social-Media-Deepfake-Kampagnen dürften dabei regelmässig nur den ressourcenstärksten Akteuren gelingen.

Die Diskussion der Deepfake-Szenarien hat Implikationen für mögliche Massnahmen zum Schutz vor Deepfakes und zur Schadensbegrenzung. Ein Anknüpfungspunkt ist das Ziel: Deepfakes, die gegen einzelne Personen, Gruppen oder Institutionen gerichtet sind, verdeutlichen die Notwendigkeit von Massnahmen durch die betroffenen Personen, Gruppen oder Institutionen selbst. Deepfakes, die auf die Erreichung eines grossen Adressatenkreises abzielen, verdeutlichen hingegen die Notwendigkeit der gesellschaftlichen Sensibilisierung und Resilienz, damit ein Deepfake erst gar nicht die intendierte Wirkung entfalten kann. Diese und weitere mögliche Massnahmen diskutieren wir ausführlicher in Abschnitt 7.4.

**Tabelle 9:** Szenarien zu Einsatzmöglichkeiten von Deepfakes in der Politik (eigene Zusammenstellung)

| Ebene | Szenario | Übergeordneter Effekt | Angriffstyp | Adressatenkreis |
|---|---|---|---|---|
| Individuum | 1 | Erpressung bzw. Einschüchterung eines Politikers | Gefälschte private Aussagen oder Handlungen | Enger Adressatenkreis |
| | 2 | Rufschädigung eines Politikers | Gefälschte private Aussagen oder Handlungen | Eher enger bis weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | 3 | Anstachelung zu Gewalttaten gegen einzelne Politikerinnen und Politiker | Gefälschte private Aussagen oder Handlungen | Enger bis weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| Institutionen | 4 | Rufschädigung einer Partei oder (staatlichen) Institution | Gefälschte private Aussagen oder Handlungen | Eher enger bis weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | 5 | Erbeutung von vertraulichen Informationen | Social-Engineering-Angriff | Enger Adressatenkreis |
| | | | Überwindung von Sicherheitsmassnahmen | |
| | | | Fälschung von privaten Aussagen und Handlungen | |

| Ebene | Szenario | Übergeordneter Effekt | Angriffstyp | Adressatenkreis |
|---|---|---|---|---|
| Gesellschaft | 6 | Beeinflussung einer Wahl | Gefälschte private Aussagen oder Handlungen | Eher enger bis weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | 7 | Beeinflussung von politischen Entscheidungsprozessen | Synthetische Social Botnets | Enger bis weiter Adressatenkreis |
| | | | Gefälschte private Aussagen oder Handlungen | |
| | 8 | Verschärfung sozialer Spannungen | Gefälschte private Aussagen oder Handlungen | Weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | 9 | Beschädigung des demokratischen Wesens | Gefälschte private Aussagen oder Handlungen | Weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | 10 | Gefährdung der öffentlichen Sicherheit | Gefälschte private Aussagen oder Handlungen | Enger bis weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | | | Überwindung von Sicherheitsmassnahmen | |
| | | | Social-Engineering-Angriff | |
| | 11 | Beeinflussung der internationalen Beziehungen | Gefälschte private Aussagen oder Handlungen | Weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | | | Überwindung von Sicherheitsmassnahmen | |
| | | | Social-Engineering-Angriff | |

## 6.4. Zwischenfazit

Die wissenschaftliche Literatur zur Verwendung von Deepfakes in politischen Kontexten beschäftigt sich mit zahlreichen möglichen Auswirkungen von Deepfakes auf die Politik. Allerdings diskutieren unterschiedliche Texte unterschiedliche Gefährdungsaspekte in je verschiedener Weise: Es mangelt also einer klaren Systematik. Zudem besteht ein Fokus insb. auf die Gefährdung der Demokratie insgesamt und die Manipulation von demokratischen Wahlen mittels Deepfakes. Insofern fehlt ein systematischer Gesamtüberblick der durch Deepfakes hervorgerufenen Herausforderungen für die Politik. Sofern Szenarien diskutiert werden, beruhen diese nicht auf einer stringenten Struktur. Arbeiten mit Bezug zu Schweizer Politik fehlen hierbei gänzlich.

Zur Adressierung der offenen Punkte wurde zunächst eine Umfrage unter Schweizer Parlamentariern und Verwaltungsmitarbeitenden durchgeführt. Im Anschluss wurden auf Basis des Literaturüberblicks und der Umfrageergebnisse ausführliche Szenarien erarbeitet, welche Rolle Deepfakes in der Schweizer Politik spielen könnten. Schliesslich wurden zu den einzelnen Szenarien mögliche Massnahmen zum Schutz und zur Schadensbegrenzung zugeordnet.

**Umfrage unter Schweizer Parlamentarierinnen und Parlamentariern**

In der Umfrage gab eine Mehrheit der Befragten an, dass Deepfakes entweder bereits ein Diskussionsthema im politischen Betrieb sind oder sie äusserten den Wunsch, dass dem Thema mehr Aufmerksamkeit gewidmet werden sollte. Bei den wahrgenommenen Chancen und Risiken zeigte sich ein eindeutiges Bild: Die Befragten sahen fast ausschliesslich Risiken. Bedenken fokussierten sich insbesondere auf die Gefahren für die Schweizer Demokratie und die politischen Institutionen, die von den Teilnehmenden vertreten werden. Hinsichtlich beider Aspekte sahen die Befragten zudem mehrheitlich eine hohe oder sehr hohe Eintrittswahrscheinlichkeit. Das Risiko, dass ein Deepfake über einen selbst kursiert bzw. dass man mittels eines Deepfakes getäuscht wird sowie die aus Deepfakes resultierenden Risiken für internationale Beziehungen wurden ebenfalls häufig als relevantes Risiko benannt. Die Mehrheit der Teilnehmenden war allerdings zugleich der Ansicht, dass die Wahrscheinlichkeit des Eintritts derartiger Risiken als eher niedrig bis sehr niedrig einzustufen ist. Die Befragung zeigt schliesslich auch eine Tendenz, dass gegenwärtig noch zu selten konkrete Schutzmassnahmen gegen Deepfakes ergriffen werden.

### Szenarien zu Deepfakes in der Politik

Zur Konkretisierung der möglichen Nutzung von Deepfakes in der Schweizer Politik wurden Szenarien erarbeitet. Diese machen systematische Aussagen über mögliche Angreifer bzw. Verursacher, den Angriffstyp sowie den Adressatenkreis. Die Szenarien sind insbesondere für die spätere Diskussion möglicher Massnahmen zum Schutz und zur Schadensbegrenzung relevant (vgl. Kapitel 7.4). Darüber hinaus können die Szenarien Politikerinnen und Politiker, Parteien und andere Akteure bei der Erstellung eines eigenen Risikoprofils unterstützen.

# 7. Deepfakes in der Wirtschaft

*Murat Karaboga, Greta Runge & Michael Friedewald*

Studien weisen darauf hin, dass Deepfakes seitens Einzelpersonen oder organisierter Akteure zu Zwecken wie Ausspähung von Geschäftsgeheimnissen, Erpressung, Einschüchterung oder übler Nachrede eingesetzt werden. Es wird erwartet, dass Deepfakes sich in das bestehende Repertoire von Cyberangriffen einreihen werden und neue Angriffsmöglichkeiten eröffnen (Bateman 2020; Europol 2020; ENISA 2023). Zugleich wird Deepfake-Technologien zugerechnet, auch wirtschaftliche Chancen mit sich zu bringen, die durch Unternehmen ergriffen werden könnten (Kalpokas/Kalpokiene 2022: 55 ff.). Im Vergleich zur Erforschung der Effekte von Deepfakes auf die Bevölkerung sowie der möglichen politischen Auswirkungen ist die Diskussion der potenziellen wirtschaftlichen Konsequenzen von Deepfakes im Frühstadium.

Angesichts der Erwartung, dass erwünschte wie unerwünschte Effekte von Deepfakes auf die Wirtschaft zunehmen werden, und der Feststellung, dass es sich dabei um ein bislang eher vernachlässigtes Thema handelt, verfolgt das vorliegende Kapitel das Ziel, die gegenwärtige und künftige Rolle von Deepfakes in der Wirtschaft zu untersuchen. Die konkreten Forschungsfragen lauten:

**FF 6.1:** Welche Herausforderungen durch Deepfakes sind im Hinblick auf Unternehmen zu erwarten?

**FF 6.2:** Welche Chancen bringen Deepfake-Technologien mit sich?

**FF 6.3:** Wie können sich Unternehmen vor Deepfakes schützen?

Im Folgenden werden zunächst (vgl. Kapitel 7.1) auf Grundlage des Stands der Literatur grundlegende Herausforderungen durch Deepfakes in der Wirtschaft herausgearbeitet. Daran schliesst sich die Diskussion von Chancen durch Deepfakes (vgl. Kapitel 7.2) an, die auf einer empirischen Medieninhaltsanalyse basiert. Im Anschluss werden, analog zum «Deepfakes in der Politik»-Kapitel, die Ergebnisse von neun Szenarien in Form einer Kurzzusammenfassung vorgestellt (vgl. Kapitel 7.3), wie Deepfakes konkret im Bereich der Wirtschaft in Schädigungsabsicht eingesetzt werden könnten. Zuletzt werden, basierend auf den Szenarien zu Politik und Wirtschaft, mögliche Massnahmen zum Schutz und zur Schadensbegrenzung (vgl. Kapitel 7.4) aufgezeigt.

## 7.1. Herausforderungen von Deepfakes in der Wirtschaft

**Einsatzmöglichkeiten von Deepfakes für wirtschaftliche Ziele mit Schädigungsabsicht**

Instruktive Vorarbeiten, die einen Überblick zum Einsatz von Deepfake-Angriffen gegen Unternehmen und die Wirtschaft bieten, stammen von Europol (2020) und der Carnegie Endowment for International Peace (Bateman 2020). Beide Studien kommen in ihrer jeweiligen Analyse der Einsatzmöglichkeiten von Deepfakes in schädigender Gewinnerzielungsabsicht zu überlappenden Schlüssen: Identitätsdiebstahl, Identitätsbetrug und die Initiierung betrügerischer Zahlungen seien etwa möglich durch Social-Engineering-Angriffe mittels Deepfake-E-Mail- und *Voice-Phishing*, die Fälschung von Onlineidentitäten bzw. Dokumentenbetrug und Täuschung von Know-Your-Customer-Mechanismen (ebd.: 8; Europol 2020: 52). Sprachbasierte Betrugsmethoden gelten deshalb als besonders gefährlich, weil sie Social Engineering betreiben, also die menschliche Fehlbarkeit ausnutzen und so moderne technische Sicherheitsmassnahmen wie Antiviren-Scans oder VPNs umgehen können (Brode 2020). Das FBI antizipiert insb. sehr glaubwürdige Deepfake-basierte Spearphishing- und ausgeklügelte Social-Engineering-Angriffe (FBI 2021). Obwohl ChatGPT dies explizit verhindern soll, wurde dessen Sperre bereits umgangen und das Werkzeug bereits zur Generierung von Malware-Code und von Texten für Phishing-E-Mails missbraucht, die zur weiteren Automatisierung und Vereinfachung von Cyberattacken verwendet werden können (Goodin 2023). Deepfakes könnten auch zum Zwecke des Identitätsbetrugs für Bewerbungen auf Remote-Work-Stellen verwendet werden (FBI 2022; Coldewey 2022).

Mit den Angriffstechniken gefälschter privater Aussagen und synthetischer Social Botnets sei die Beeinflussung der öffentlichen Meinung möglich, um eine Störung und Manipulation von Aktienmärkten und -kursen sowie die Initiierung von Panikkäufen und Bankenruns zu bewirken. Gefälschte private Aussagen hätten bei diesen Szenarien weniger die Schädigung der im Deepfake abgebildeten Person zum Ziel als vielmehr der Institutionen bzw. des Unternehmens, die sie vertritt und ggf. des Wirtschaftssektors oder gar des Wirtschaftssystems (ebd.). Darüber hinaus könnten Deepfakes zur Erpressung von Geld, Informationen oder Zugängen eingesetzt werden (Bateman 2020: 13 f.; Europol 2020: 52 ff.) und zum Zwecke digitalen Astroturfings verwendet werden. Digitales Astroturfing beschreibt die Vortäuschung der Existenz einer Graswurzelbewegung bzw. eines öffentlichen Meinungsbildes zu einem Thema. Laut Bateman (2020:

13f.) könnten KI-generierte Textinhalte dazu genutzt werden, öffentliche Konsultationen zu Finanzregulierungen mit gefälschten Positionierungen zu überfluten, um den Gesetzgebungsprozess in Richtung einer gewünschten Position zu manipulieren.

Angriffe mittels synthetischer Social Botnets basieren darauf, dass manipulierte oder synthetisch generierte Audio-, Bild-, Video- und Textinhalte dazu genutzt werden, gefälschte Social-Media-Profile zu erstellen und zu betreiben. Mittels Gesichtsgeneratoren können beispielsweise synthetische Gesichter erstellt werden, die schwieriger zu erkennen sind als herkömmliche Bots, die sich geklauter Fotos real existierender Personen bedienen (Bateman 2020: 7). Durch die Ausschöpfung der Möglichkeiten moderner Chatbots könnten die Social Bots miteinander und anderen echten Social-Media-Nutzenden so kommunizieren, dass sie nicht oder nur schwer als Bots erkannt werden. Dabei könnten die Social Bots unterschiedliche Bot-Charaktere besitzen und konsequent in ihrer Rolle verbleiben. Dies würde die Glaubwürdigkeit der Bots erhöhen und damit ihre Detektion erschweren. Ebenso könnten die Bots auf diese Weise in ganz unterschiedliche Communities hineinwirken, sodass die Wirkung eines Deepfake-Angriffs durch höchstmögliche Streuung maximiert wird. Sobald der Aktivierungsbefehl kommt, könnten sie dann auf koordinierte Weise Postings mit dem Deepfake absetzen, während sie einerseits in ihrer zuvor definierten Charakterrolle verbleiben und andererseits klassische Botnet-Angriffsdetektoren umgehen, weil nicht alle Postings zeitgleich abgesetzt würden, sondern mit ausreichend viel zeitlichem Abstand (van Huijstee u.a. 2021). Aus dem Bereich der generellen Literatur rund um den «betrügerischen Einsatz» neuer Technologien findet sich etwa auch der Werbebetrug in Form des sog. Fake-Influencer-Marketings. Dabei werden gefälschte Social-Media-Profile zur Generierung von Profit durch Werbeeinnahmen verwendet (Cavazos 2019). Die Möglichkeit, solche Fake-Profile durch Deepfake-Technologien glaubwürdiger erscheinen zu lassen, könnte auch die Effektivität von Werbebetrugsmaschen gesteigert werden.

Andere Arbeiten (Wojewidka 2020; Sidelov 2022) diskutieren die Gefahr, die von Deepfakes für biometrische Systeme ausgeht. Demnach seien Deepfakes dazu geeignet, Identitätsbetrug zu begünstigen, indem Gesichts- oder Stimmerkennungssyteme überlistet werden.

Die Mehrheit der vorhandenen Analysen zu Deepfakes liefert keine über die in den o.g. Arbeiten hinausgehenden Erkenntnisse, weil sie sich auf die Gefahren für Politik und Demokratie fokussieren (z.B. Chesney/Citron 2018; Westerlund 2019; Marchetti 2022). Einig ist man sich in der Literatur jedenfalls darin, dass sich Deepfakes in das Repertoire bestehender Cyberangriffe einfügen werden.

Daher sollen heutige Formen von Cyberangriffen im Folgenden näher betrachtet werden.

**Überblick zu Cybergefahren**

Deepfake-basierte Aktivitäten im Bereich der Wirtschaftskriminalität stellen keine vollständig neuen Gefährdungen dar, sondern fügen sich zumeist in das Repertoire bekannter Cyberoperationen ein. Im Bereich der Cyberkriminalität gelten Ransomware-Attacken (auch bekannt als Erpressungs-Schadsoftware-Attacke) derzeit als Dreh- und Angelpunkt (Symantec Threat Hunter Team 2021). Bei einer Ransomware-Attacke ist es das Ziel, Schadsoftware in ein Computersystem einzuschleusen, das System zu verschlüsseln und für die Entschlüsselung eine Lösegeldsumme zu erpressen. Dies erfolgt zumeist über sog. Social-Engineering-Techniken, die als Angriffsvektor dienen. Dabei werden durch einen Angreifer menschliche Eigenschaften wie Hilfsbereitschaft, Vertrauen, Angst oder Respekt vor Autorität ausgenutzt, um Personen bzw. Mitarbeitende zu manipulieren. Die häufigste Variante einer solchen Manipulation erfolgt mittels gefälschter E-Mails, auch Phishing-Mails genannt. Dabei werden die Opfer von den Angreifern mittels zunehmend authentisch erscheinender E-Mail-Betreffzeilen und -Inhalte getäuscht, damit sie angehängte «Köder»-Dokumente öffnen und den enthaltenen Schadcode ausführen. Dadurch wird dem Angreifer die Infiltration nicht nur des einzelnen Computers, sondern auch des Computernetzwerks ermöglicht. Diese Infiltration hat zunächst zum Ziel, die administrative Kontrolle über das gesamte Computersystem und -netzwerk zu erhalten. Neben dem Angriffsvektor «Social Engineering» setzen Angreifer auch auf sog. *Drive-By-Infektionen* mittels *Exploit-Kits* sowie die Ausnutzung von Schwachstellen in Servern und ungeschützte Fernzugänge (BSI 2022b: 5 f.).

In der Vergangenheit wurde das jeweilige System dann seitens der Angreifer verschlüsselt, um die Opfer zur Zahlung einer Lösegeldsumme zu erpressen, damit sie wieder Zugang zum System bzw. die Kontrolle darüber zurückerhalten. In den vergangenen Jahren wurde zusätzlich die Exfiltration von Daten aus dem infiltrierten System zunehmend verbreiteter, um Profite zu steigern. Im Falle von sensiblen und wertvollen Kunden- und Unternehmensdaten erhalten Angreiferinnen und Angreifer so einen Hebel, um ihre Opfer aus Furcht vor der Datenpreisgabe auch dann zur Lösegeldzahlung zu erpressen, wenn diese ansonsten mittels Backups ein Reset ihres infiltrierten Systems durchführen könnten. Üblicherweise wird in diesen Fällen zur Erhöhung des Drucks auf die Opfer ein Teil der Daten veröffentlicht, um der Drohung Nachdruck zu verleihen. In der Regel erhalten die Opfer nach erfolgter Lösegeldzahlung wieder die Kontrolle über das Computersystem bzw. ihre Daten zurück. Es hat aber auch schon Fälle gegeben

(Ordinypt, NotPetya und GermanWiper), in denen Lösegeldforderungen gestellt, aber die Daten dennoch unwiderruflich zerstört wurden. Diese Angriffe zeigen, dass die Motivation der Angreiferinnen und Angreifer auf finanziellen Gewinn und Sabotage zugleich oder nur auf Sabotage zielen kann (BSI 2022b).

Durch eine Ransomware-Attacke können Eigen-, Reputations- sowie Fremdschäden entstehen. Zu Eigenschäden zählen etwa die unmittelbaren Kosten durch Betriebsbeeinträchtigungen bzw. -unterbrechungen wie ein Produktionsstillstand. Weitere Schäden können durch Krisenreaktionsmassnahmen durch Mitarbeitende oder externe Fachleute, für Forensik und Wiederherstellung sowie durch gesetzliche Vorgaben (Aufsichtsbehörden- und Betroffenenunterrichtung oder Bussgelder) entstehen. Folgeschäden können durch den Abfluss von wichtigen Kunden- und Unternehmensdaten an die Konkurrenz entstehen. Reputationsschäden können entstehen, wenn das Vertrauen in ein Unternehmen bzw. eine Organisation infolge eines Angriffs sinkt, Kunden abwandern und ggf. Aktienkurse fallen. Weitere Kosten können durch Investitionen in den neuen Reputationsaufbau mittels Werbung, Kundenbindungsmassnahmen und Imagekampagnen entstehen. Fremdschäden können entstehen, wenn gesetzliche, vertragliche oder sonstige Verpflichtungen gegenüber Dritten infolge eines Angriffs nicht eingehalten werden können (BSI 2022a). Wie der Angriff auf den US-amerikanischen Pipelinebetreiber Colonial Pipeline im Jahr 2021 demonstrierte, können insbesondere Angriffe auf kritische Infrastrukturen weitreichende Fremdschäden nach sich ziehen. Trotz der Zahlung des Lösegelds führten Verzögerungen bei der Datenentschlüsselung zu Treibstoffknappheit, Preissteigerungen und Panikkäufen, sodass ein regionaler Notstand ausgerufen wurde (Symantec Threat Hunter Team 2021). Grundsätzlich können sich derartige Angriffe gegen beliebige Ziele richten, als besonders relevant gelten gemäss der nationalen Strategie zum Schutz kritischer Infrastrukturen (SKI) die Energieversorgung, der Verkehr, die Lebensmittelversorgung, Finanzdienstleistungen und die Telekommunikation. Bekannte Fälle aus der jüngeren Vergangenheit in der Schweiz sind die Ransomware-Angriffe Anfang 2023 gegen die SBB (Mäder 2023) und die NZZ (NZZ 2023).

Eine genaue Quantifizierung der durch Ransomware-Attacken verursachten wirtschaftlichen Schäden ist nicht möglich, weil eine entsprechende Datenbasis fehlt. Die meisten betroffenen Organisationen ziehen es vor, Stillschweigen insb. über erfolgreiche Angriffe zu bewahren, um einen Reputationsschaden zu verhindern. Unter anderem zur Verbesserung der Informationslage über Cyberangriffe wurden in den vergangenen Jahren in den USA und der EU Gesetze verabschiedet, die Betreibende von kritischen Infrastrukturen dazu verpflichten,

Cyberangriffe auf ihre Systeme innerhalb von 24 bis 72 Stunden an zuständige Behörden zu melden. Eine vergleichbare Meldepflicht für Betreiber von kritischen Infrastrukturen bei Cyberangriffen wurde Mitte 2023 auch in der Schweiz eingeführt (Jaun/Züllig 2023). Darüber hinaus melden schon heute viele Schweizer Unternehmen (aber auch Privatpersonen) Fälle von Cyberkriminalität (darunter sowohl erfolgreiche Angriffe als auch gescheiterte Angriffsversuche) auf freiwilliger Basis an das BACS (Bundesamt für Cybersicherheit, bis Ende 2023 NCSC) (Bundesrat 2023b).

**Angreifer**

Angreiferinnen und Angreifer sind einerseits kriminelle Akteure, die eher ein ökonomisches Interesse haben, und andererseits staatliche Akteure, die sowohl wirtschaftliche als auch politische Interessen haben können (Insikt Group 2023). Die Forschung der Insinkt Group (2023) zeigt, dass bereits ein illegaler Onlinemarkt rund um Deepfakes entstanden ist. Vertrieben werden insb. Dienstleistungen zur Produktion von Deepfake-Videos, aber auch von gefälschten Bankkarten, Dokumenten, Deepfake-Fotos usw. Für einfache Deepfake-Videos betrug der Preis 20 US-Dollar pro Videominute. Neben dem Angebot an konkreten Dienstleistungen dienen die Darkweb-basierten Onlineforen auch zum allgemeinen Wissensaustausch über Deepfake-Technologien und technologische Entwicklungen. Auch finden sich Anleitungen und Lektionen zur Deepfake-Produktion und z.T. entsprechende Anwendungen zum Download (ebd.: 3).

Vor allem englisch- und russischsprachige Darknetseiten bzw. -foren gelten als Hauptquellen für die Diskussion, die Bewerbung, den Vertrieb und Kauf von Deepfake-bezogenen Produkten und Diensten. Auch türkisch-, spanisch- und chinesischsprachige Darknetseiten und -foren seien häufig anzutreffen (Insikt Group 2023: 3). Im Hinblick auf gegen westliche Staaten gerichtete Cyberangriffe ist häufig von chinesischen (Meiritz u.a. 2021), russischen (Baumann u.a. 2023) und nordkoreanischen Angreifern die Rede (Der Spiegel 2023). Konkret gegen Schweizer Organisationen gerichtete Cyberangriffe sind Erhebungen des Council on Foreign Relations zufolge vor allem auf chinesische, russische und iranische Urheber zurückzuführen (Council on Foreign Relations 2023). Die Attribution der Urheber eines Angriffs ist im Bereich der Cyberkriminalität jedoch äusserst schwierig, sodass derartige Angaben mit Vorsicht zu geniessen sind. Auf eine entsprechende Anfrage unsererseits antwortete das fedpol, dass bei Betrügereien Kryptowährungen betreffend häufig nordkoreanische Akteure beteiligt seien. Ransomware-Angriffe gingen grösstenteils von russischsprachigen und in geringerem Masse von chinesischen Gruppen aus. Romance-scams würden oft von Gruppen aus Westafrika durchgeführt und bei Onlineinvestitionsbetrügereien seien zahlreiche Gruppen involviert. Kleinere Betrugsfälle auf Online-

marktplätzen oder im Immobilienbereich würden meist von Einzelpersonen in der Schweiz begangen. Diese Trends seien jedoch nicht spezifisch für die Schweiz.

Wie im Falle der russischen Cyberkriegstruppe *Vulkan* bekannt wurde, gehen die Angreifer systematisch vor und durchsuchen das Internet grossflächig nach potenziellen Schwachstellen, über die sie in fremde Systeme eindringen und Schaden anrichten können. In den Schulungsunterlagen der Vulkan-Gruppe war etwa auch das Schweizer Aussenministerium als mögliches Angriffsziel benannt (Baumann u.a. 2023).

Digitale Umgebungen gelten grundsätzlich als sehr anfällig für Cyberangriffe bzw. als einladend für Angreifer. Diese Anfälligkeit lässt sich mit der *Routine Activitiy Theory* (RAT) gut veranschaulichen: Demnach sind Angreifende keine blossen Opportunisten, sondern bei der Auswahl ihrer Ziele von rationalen Motiven getrieben. Neben der individuellen Motivation einer Täterin oder eines Täters seien ausserdem die Spezifika des Ziels und das Fehlen eines ausreichenden Schutzes entscheidend (Yar 2005; Leukfeldt/Yar 2016). Interessant für die Diskussion um Deepfakes, in denen häufig Personen des öffentlichen Lebens dargestellt werden, ist die Frage, wann ein Ziel als geeignet gilt. Laut RAT sind vier Kriterien entscheidend: Wert, Grösse/Trägheit, Sichtbarkeit und Zugänglichkeit.

Der Wert eines Deepfake-Angriffs bemesse sich etwa daran, wie viel Befriedigung aus der Erschaffung von Deepfake-Pornografie gezogen werde oder welcher materielle Gewinn aus einer Deepfake-Erpressung erzielt werden könne. Trägheit sei – anders als etwa beim Diebstahl eines physischen Objekts, das vom Ort des Geschehens unauffällig wegbewegt werden muss – im Bereich von Deepfake-Angriffen praktisch nicht vorhanden, weswegen die Angriffshürde sinke. Die Sichtbarkeit einer Person oder eines Unternehmens erhöhe wiederum die Wahrscheinlichkeit dafür, als Ziel ausgewählt zu werden. Die weite Zurschaustellung alltäglicher Lebenspraxen mittels sozialer Medien erhöhe somit grundsätzlich die Gefahr, als Ziel ausgewählt zu werden. Die Zugänglichkeit bezieht sich schliesslich auf die Kontextbedingungen, die es Angreifenden einfacher oder schwerer machen, ein Zielobjekt zu erreichen. Im Bereich von Deepfakes bezieht sich dieser Aspekt vor allem auf den Zugang zu Trainingsmaterial über eine Person und zu Verbreitungskanälen zur Zirkulation des Deepfakes, die beide als hoch einzustufen sind (Kalpokas/Kalpokiene 2022: 65 f.).

**Zur Situation in der Schweiz**

Die Schweizer Wirtschaft gilt als eine der innovativsten und produktivsten Volkswirtschaften der Welt. Nach Bruttoinlandsprodukt pro Kopf lag die Schweiz 2021 laut IWF auf dem weltweit dritten Platz. Obwohl die Schweiz gemessen an der Einwohnerzahl unter 194 Ländern weltweit auf dem 100. Platz liegt, nahm

die Schweizer Wirtschaft gemessen am nationalen BIP den 21. Platz ein (IMF 2022). Viele international tätige Konzerne und Unternehmen haben ihren Hauptsitz, Niederlassungen oder wichtige Datenzentren in der Schweiz. Daher gelten Schweizer Unternehmen als attraktives Ziel von Cyberangriffen (Bundesrat 2023b).

Trotz der bedeutenden Rolle der Schweizer Wirtschaft gilt die Schweizer Cybersicherheitsarchitektur als vergleichsweise ausbaufähig: Im Global Cybersecurity Index 2020 der Internationalen Fernmeldeunion (ITU) belegte die Schweiz unter 182 Staaten den 42. Platz im Hinblick auf das Level der nationalen Cybersicherheit (ITU 2021: 25).

In einer im Auftrag des Nachrichtendienstes des Bundes (NDB) angefertigten und im Jahr 2020 veröffentlichten Studie der Universität Bern wurde der Stand der Wirtschaftsspionage in der Schweiz untersucht. Von den befragten Unternehmen gaben 15 % an, bereits von Wirtschaftsspionage betroffen gewesen zu sein. Dabei spielte es keine Rolle, ob es sich um KMU oder Grossunternehmen handelt. Betroffen waren insbesondere die Branchen Maschinenbau und Industrie sowie Pharma und Life Science, ausserdem auch Baugewerbe/Bau, Information, Kommunikation und Verlagswesen, Luft- und Raumfahrttechnik, Rüstungsindustrie, Elektronik sowie Messtechnik. Zudem gelten der Bankensektor, Versicherungen und Immobiliensektor als potenziell gefährdet (Zwahlen u.a. 2020).

Auffällig ist, dass nur 13 % der betroffenen Unternehmen die Vorfälle bei Polizeibehörden oder Staatsanwaltschaften meldeten. Häufiger ergriffen die betroffenen Unternehmen firmeninterne Massnahmen (20 %) oder arbeiteten den Vorfall mit externer Unterstützung (16 %) auf. Mit 47 % blieb die grosse Mehrheit der Unternehmen allerdings nach einem Vorfall untätig, etwa weil sie zu wenige Anhaltspunkte über den Vorfall verfügten (Zwahlen u.a. 2020: 26).

Die NDB-Studie verdeutlicht die anhaltend hohe Relevanz (39 % aller Angriffe) von Angriffsversuchen mittels E-Mails mit kritischem Inhalt, also insb. das versuchte Einschleusen von Ransomware. Dies zeigt sich auch bei den konkreten Angriffen, wonach Phishing- und Spearfishing-Angriffe (18 %) nach Cyberangriffen auf Firmennetzwerke (19 %), also z.B. Hacking-Angriffe auf Schwachstellen, zu den häufigsten Angriffsformen zählen (Zwahlen u.a. 2020: 21 f.).

Betroffene Unternehmen gaben mit 18 % am häufigsten an, dass sie nach einem Vorfall immaterielle Schäden durch den Verlust von Wettbewerbsvorteilen davontrugen. Auch der Ausfall von Informatik (14 %) und Kunden- und Auftragsverluste (11 %) wurden genannt. Reputationsverluste oder negative Presse

folgten mit 8 % und Kosten für Rechtsstreitigkeiten mit 7 %. Immerhin 15 % der betroffenen Unternehmen gaben aber zugleich an, dass kein Schaden entstanden oder der Schaden unbekannt sei. Derartige Schäden können weitreichende Konsequenzen nach sich ziehen: In den meisten Fällen zogen die Angriffe zwar keine messbaren Konsequenzen nach (38 %) sich oder konnten innerhalb von ein paar Stunden (20 %) behoben werden. Doch 11 % der betroffenen Unternehmen gaben an, dass der Angriff die Existenz des Unternehmens gefährdet habe (Zwahlen u.a. 2020: 27).

Der Grossteil der befragten Unternehmen stützt sich im Wesentlichen auf firmeninterne Massnahmen zur Prävention von Angriffen. Auf staatliche Unterstützung oder externe Beratung wird nur selten zurückgegriffen. Typische firmeninterne Schutzmassnahmen sind die Regelung und Einschränkung von Zugriffsrechten der Mitarbeitenden auf Dokumente und Daten (12 %), die Vernichtung vertraulicher Akten und Datenträger (8 %) und die Klassifizierung von Dokumenten und Daten (7 %). Die getroffenen Schutzmassnahmen richten sich sowohl gegen Wirtschafts- als auch Industriespionage und können in fünf Kategorien unterteilt werden (Zwahlen u.a. 2020: 31 ff.).

- Strukturelle/organisatorische Massnahmen
    - Bspw. eine gut ausgestattete IT-Abteilung oder die Einsetzung umfangreicher Sicherheitskonzepte
- Personelle Massnahmen
    - Schulung und Sensibilisierung der Mitarbeitenden, sowohl informelle, mündliche Formen als auch regelmässige Schulungen, teils in Kooperation mit dem NDB
    - Background-Checks von Bewerberinnen und Bewerbern, ggf. auch Personensicherheitsprüfungen (PSP) des Eidgenössischen Departements für Verteidigung, Bevölkerungsschutz und Sport (VBS)
    - Erhöhung der Zufriedenheit von Mitarbeitenden zur Steigerung ihrer Vorsichtigkeit und Verringerung der Wahrscheinlichkeit, nach einem Arbeitgeberwechsel gegen den vorherigen Arbeitgeber aktiv zu werden
- Konkrete IT-Massnahmen wie Instrumente und Technologien
    - Zugriffsrechtesystem auf (sensible) Daten und Dienste
    - Firewalls, Spamfilter und Virenschutzsoftware

  - Nicht mit dem Internet verbundene, eigene Server zur Lagerung sensibler Daten
  - Externe Backup-Lösungen
  - Ggf. Cloud-Lösungen
- Physische/technische Massnahmen
  - Dazu zählen insb. physische Zugangskontrollmassnahmen
- Konkreter Schutz von Produkten und Know-how
  - Patentierung und Herausgabe von lückenhaften Produkten, um Reverse Engineering zu erschweren

## 7.2. Chancen von Deepfakes in der Wirtschaft

Wie die bisherigen Erörterungen gezeigt haben, waren die mit Deepfakes assoziierten Herausforderungen in den vergangenen Jahren Gegenstand zahlreicher Artikel und Studien. In der einschlägigen Literatur ist hingegen selten bis gar nicht Thema, welche Chancen sich durch Deepfakes ergeben könnten (Godulla u.a. 2021). Zur Untersuchung allfälliger Chancen führten wir eine empirische, medieninhaltsanalytische Erhebung der Nutzung von Deepfake-Technologien für erwünschte Zwecke durch. Dabei zeigt sich, dass Chancen durchaus vorhanden sind, diese allerdings auch mit neuen Herausforderungen einhergehen können. Diese Ergebnisse werden im Folgenden vorgestellt.

### 7.2.1. Methodisches Vorgehen

Zur Untersuchung der Frage, welche Chancen Deepfake-Technologien mit sich bringen, führten wir eine Medieninhaltsanalyse durch. Diese stützte sich v.a. auf die Analyse von nationalen Tageszeitungen, internationalen Medien und deutschen IT-Fachmedien, weil diese ein grosses Spektrum an gesellschaftlichen Debatten und der öffentlichen Meinung abdecken (Mayring 2010). Mit der Medienauswahl wurde eine Mischung aus eher spekulativeren Perspektiven (in IT-Fachmedien) und heute real anzutreffenden Deepfake-Anwendungen (v.a. in Tageszeitungen und internationalen Medien) untersucht. Wo passend, wurde allerdings auch auf Eindrücke aus der allgemeinen Literaturrecherche zu Deepfake-Technologien (vgl. Kapitel 2) zurückgegriffen. Dabei wurde darauf

geachtet, dass nationale und internationale Medien, Print- und Onlineformate sowie die Abbildung verschiedener Meinungsspektren Berücksichtigung finden. Das Untersuchungsmaterial wurde zum einen über das Onlinerecherchetool für Presseerzeugnisse *LexisNexis* und zum anderen händisch über die Webseiten der betreffenden News-Portale erhoben. Als Untersuchungszeitraum wurde das Erscheinungsjahr des ersten Deepfakes 01.01.2017 bis 11.08.2022 festgelegt. Insgesamt wurden 445 Artikel erhoben, die sich direkt auf das Untersuchungsthema *Deepfakes* beziehen.[287]

Anschliessend wurden die Artikel daraufhin gesichtet, ob *Anwendungsfelder* oder *-bereiche* von Deepfakes sowie Chancen und Risiken der Technologie in den Artikeln benannt und diskutiert werden. Nach dem kategoriengeleiteten Selektionsschritt verdichtete sich das Untersuchungsmaterial auf insgesamt 241 Artikel. 38 Artikel wurden in den Schweizer Medien und 131 Artikel in internationalen Zeitungen und Zeitschriften identifiziert. In den deutschsprachigen IT-Fachmedien konnten insgesamt 72 relevante Artikel erhoben werden.

**Tabelle 10:** Untersuchungsmaterial Medieninhaltsanalyse (Druckauflagen laut Stand 2022)

| Medium | Gesamtzahl der Beiträge | Ausgewählte Beiträge | Druckauflage |
|---|---|---|---|
| **Tageszeitungen Schweiz** | | | |
| Tages-Anzeiger | 25 | 6 | 131.000 (2019) |
| Berner Zeitung | 1 | 1 | 111.014 (2021) |
| Luzerner Zeitung | 15 | 8 | 110.000 (2019) |
| St. Gallen Tageblatt | 11 | 4 | 109.000 (2019) |
| Neue Zürcher Zeitung | 33 | 16 | 67 458 (2021) |
| SRF | 9 | 3 | |
| **IT-Fachmedien (deutschsprachig)** | | | |
| Heise online | 61 | 41 | – |
| Golem.de | 12 | 6 | – |
| Computerbild.de | 8 | 4 | – |
| T3n Magazin | 53 | 21 | 48.850 (2019) |

[287] Im Rahmen der Erhebung wurden verschiedene Schreibweisen von Deepfakes *(Deepfake(s), Deep Fake(s), Deep-Fake(s)* sowie das Synonym *synthethische Medien* berücksichtigt.

| Medium | Gesamtzahl der Beiträge | Ausgewählte Beiträge | Druckauflage |
|---|---|---|---|
| **Internationale Zeitungen und Zeitschriften** | | | |
| New York Times | 177 | 96 | 1.163.000 (2021) |
| USA Today | 14 | 10 | 1.315.780 (2021) |
| Peoples Daily | 10 | 7 | 2.500.000 (2022) |
| Spiegel (Online) | 25 | 18 | 731.692 Druck-exemplare und 245.326 Digital-Abos (2022) |
| **Gesamtanzahl** | **454** | **241** | |

Mittels der qualitativen Inhaltsanalyse, einer empirischen Methode, «die systematisch erfahrungsbasierte Erkenntnisse generiert» (Baur/Blasius 2019), wurden die Artikel anschliessend analysiert. Dabei wurden induktive Kategorien aus dem Material heraus formuliert (Mayring 2010). Insgesamt wurden drei Anwendungsfelder identifiziert: *Unterhaltung, Bildung* sowie *Werbung.*

Neben der inhaltsanalytischen Betrachtung wurden zudem konkrete Anwendungen in Form von Software, mobilen Apps oder Webapplikationen erhoben. Insgesamt konnten 102 Applikationen zur Erstellung von Deepfakes in unterschiedlichen Bereichen, von unterschiedlichen Herstellern oder Forschenden identifiziert werden (detaillierte Liste siehe Annex A.4). Insgesamt fanden sich 78 verschiedene Hersteller, wobei zu 13 Applikationen kein Urheber ermittelt werden konnte. Die Mehrzahl der Hersteller ist in den USA ansässig, wobei für 38 Applikationen kein Herkunftsland ermittelt werden konnte. Wie in Abbildung 30 zu sehen, dominieren mobile Apps für Android/iOS (35 Applikationen), Computerprogramme (für PC/Mac/Server) (34) sowie Web-Applikationen (via Webseite) (27).

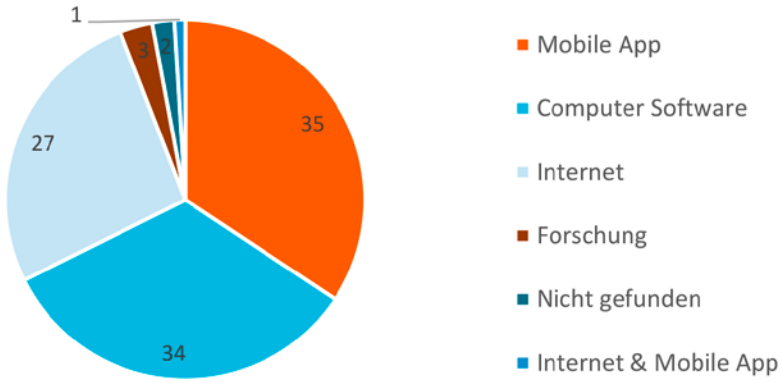


**Abbildung 30:** Anzahl der Deepfakes-Software nach verwendeter Plattform

Die meisten Anwendungen zielen auf die Erstellung synthetischer Medien/Inhalte des Gesichts ab (54), bei 30 Applikationen steht die Stimme im Vordergrund (Abbildung 31). Viele Applikationen erschienen im Jahr 2020 (12), wobei erste Software zur Manipulation von Bildmaterial und Erstellung von Videos bereits 1991 auf den Markt kam (Abbildung 32). Allerdings war das Erscheinungsdatum der Mehrzahl der Applikationen (47) nicht bestimmbar.

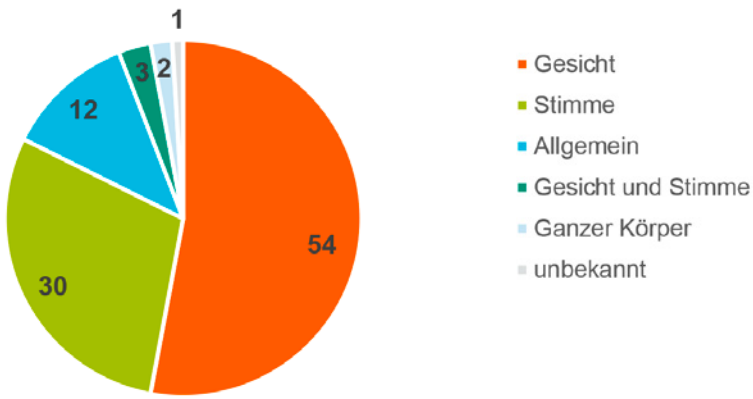


**Abbildung 31:** Anzahl der Deepfakes-Software nach Gegenstand des Fakes

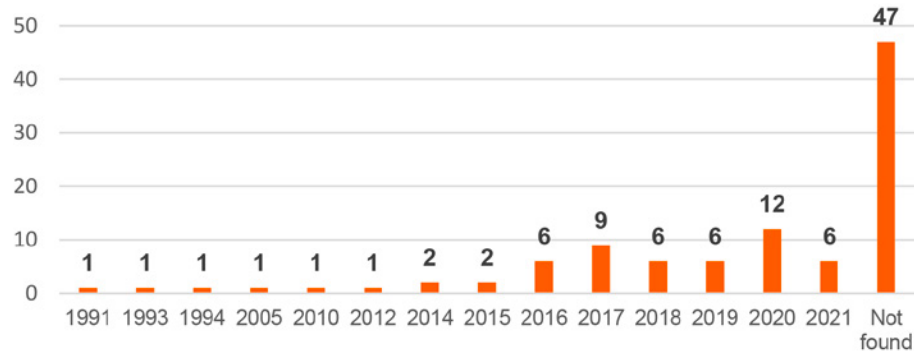


**Abbildung 32:** Anzahl der Software nach Erscheinungsjahr

### 7.2.2. Resultate

#### 7.2.2.1. Unterhaltung

Deepfake-Technologien sind schon heute vielfach im Bereich der Unterhaltungsindustrie anzutreffen. Dieser Bereich gilt auch als der zukunftsträchtigste für erwünschte Nutzungen von Deepfakes.

Dass Deepfake-Technologien vielfältige Anwendungspotenziale in der **Filmindustrie** bergen, konnte erstmals 2016 im Kinofilm «Rogue One: A Star Wars Story» erahnt werden, als der bereits 1994 verstorbene Peter Cushing mittels CGI digital «wiederbelebt» wurde. Publikum wie auch Kritikerinnen und Kritiker waren jedoch hin- und hergerissen: Einige fanden die 3D-Computergrafik beeindruckend, andere hingegen bemängelten die Qualität der unter enormen Kosten erstellten Animation (Walsh 2016). Nur wenige Jahre später wurde dieselbe Szene neu und in weitaus besserer Qualität produziert. Dem YouTuber *Shamhook*, einer Einzelperson, war es gelungen, unter Einsatz von Deepfake-Technologien der CGI-Produktion eines Hollywood-Studios nicht nur Konkurrenz zu machen, sondern nach Ansicht vieler Fans deren Arbeit auch eindeutig zu übertreffen (Chichizola 2020). Nachdem der YouTuber seine Qualitäten auch bei anderen Star-Wars-Szenen unter Beweis stellte (Aschenbrenner 2020), wurde er schliesslich von Lucasfilm eingestellt, um die Technik weiter zu entwickeln und künftig die teure CGI-Technik ggf. durch Deepfake-Technologien zu ersetzen (Bastian 2021). Inzwischen ist bei Disney ein neues Deep-Learning-basiertes Verfahren im Einsatz (Zoss u.a. 2022). Angesichts dieses Beispiels wird deut-

lich, dass mit Deepfake-Technologien nicht nur Kosteneinsparungen verbunden sind, sondern womöglich auch eine höhere Qualität bei Gesichtsanimationen, der Verjüngung oder dem Altern von Charakteren, erzielt werden kann.

Auch im Rahmen eines Dokumentarfilms (2021) über den 2018 verstorbenen Autor und Moderator Anthony Bourdain wurde Deepfake-Technologie eingesetzt, um den TV-Star mediensynthetisch auferstehen zu lassen. In diesem Fall liess ein Deepfake-Audio die Stimme von Bourdain in der Dokumentation *Roadrunner* erneut erklingen (Rosner 2021).

Insbesondere der Fall der Deepfake-Stimme von Bourdain sorgte in den USA für Diskussionen über ethische Grenzen technischer Möglichkeiten. Umstritten ist die Produktion von posthumen Deepfakes deshalb, weil es keine Möglichkeit gibt, die Zustimmung der bereits verstorbenen Person einzuholen. Im Falle von Persönlichkeiten des öffentlichen Lebens besteht ein offensichtlicher finanzieller Anreiz, ihr digitales Abbild zu erstellen, weswegen diese je nach rechtlicher Situation im entsprechenden Land durch posthume Publizitätsrechte geschützt sein können. Auch die Zustimmung oder Ablehnung durch Angehörige kann rechtlich relevant sein. Dabei kann sich aufseiten der Nachkommen allerdings ein Spannungsfeld zwischen emotionalen und finanziellen Beweggründen ergeben (vgl. zu diesem Thema die TA-SWISS-Studie «Tod im digitalen Zeitalter»).

Deepfake-Technologien können auch dafür verwendet werden, Identitäten und Minderheiten zu schützen. Dies wird am Beispiel des Films *Welcome to Chechnya* von David France deutlich, in dem Deepfakes eingesetzt worden sind, um die Verfolgung homosexueller Personen zu dokumentieren, ohne reale Menschen in Gefahr zu bringen. Jedoch wirft das Beispiel Fragen zur expressiven Sachlichkeit und dem Wahrheitsanspruch (oder Wahrheitsproduktion) von Dokumentarfilmen auf.

Im Mai 2022 wurde der verstorbene Basketballspieler Kobe Bryant Gegenstand in dem Musikvideo des US-amerikanischen Rappers Kendrick Lamar (Lamar 2022). Der Künstler setzt in seinem Video auf die Deepfake-Technologie Face Swap, sodass sich sein Gesicht im Laufe des Clips in das von anderen berühmten Persönlichkeiten verändert. Auch Teile der **Musikindustrie** sehen Chancen: Hier finden sich nicht nur auf Ebene der Musikvideos vielfältige Anwendungsmöglichkeiten. «Deepfake music» nennt sich ein Trend, dessen Software von der Firma OpenAI entwickelt wurde. Die Entwicklerinnen von OpenAI haben ihre Software mit bekannten Audio- und Textdaten gefüllt, auf deren Basis computergenerierte Songs erzeugt werden können, die beispielsweise nach Queen, Mozart oder Rammstein klingen. Das Programm zerteile die Audiosignale in

Fragmente und setze sie nach Input, Stil oder Künstler neu zusammen. Für die Musikindustrie ergeben sich hierbei viele Chancen, insb. die Senkung von Produktionskosten. Ein Beispiel hierfür ist die Abba-Avatar-Show, in der die Stars technologisch verjüngt als Hologramme auftreten. Ein anderes Beispiel ist die Veröffentlichung eines Musikvideos der K-Pop-Band Etern!ty, welches ausschliesslich aus bereits bestehendem Bildmaterial der elf Bandmitglieder entwickelt worden ist. Der Deepfake erhielt in der Fangemeinschaft positive Resonanz. Aufwendige Motion-Capturing-Verfahren könnten in Musikvideos (aber auch der Filmproduktion) durch die Produktion von Deepfakes ersetzt werden und damit Vorteile wie Kosten- und Zeiteffizienz mit sich bringen.

Die Idee, aus bestehenden Audiodaten synthetische Kommunikation zu generieren, verfolgt ein Start-up namens Neosapience aus Südkorea. Es fokussiert sich auf die Vertonung von Texten, was für die Produktion von Filmen und Musik enorme Kosteneinsparungen mit sich bringen könnte. Der potenzielle Anwendungsbereich dieser synthetischen Audios zielt unter anderem darauf, Superstars durch Stimmkopien zu imitieren oder Sprachbarrieren im interkulturellen Austausch zu beheben. Hierbei stellen sich jedoch weitreichende Fragen im Bereich der Privatsphäre und des Datenschutzes, wenn bestehendes Stimm- und Sprachmaterial zweckentfremdet wird (Chesney/Citron 2019). Auch der Bereich der Generierung von desinformativen Nachrichteninhalten könnte durch die *Sprachsynthetisierung* eine neue Dimension gewinnen. Insgesamt erscheinen jedoch je nach Anwendungsfall Fragen des Urheberrechts sowie der Qualität der KI-generierten Musik- und Stimmerzeugnisse noch offen zu sein (vgl. auch Kapitel 4.2.1.3).

In einem engen Verhältnis zu dem Einsatz von Deepfakes in der Film- und Musikindustrie stehen Deepfakes als Werkzeug für die Realisierung von **Fan-Art**. Der o.g. YouTuber *Shamook* bildet dabei nur das bekannteste Beispiel. Inzwischen sind verschiedene Anwendungen erschienen, die zu Unterhaltungszwecken *Face Swapping* bekannter Figuren aus Film und Fernsehen mit dem eigenen Gesicht erlauben. Ein anderes Beispiel für die synthetische Umsetzung von Fan-Fantasien via Deepfakes bildet die App *Wombo Dream*, bei der Lieblingscharaktere durch KI gezeichnet werden. Das Unternehmen hat in der Vergangenheit bereits die Deepfake-Lip-Sync-App *Wombo AI* veröffentlicht. Auch im **Gaming**-Bereich gibt es erste Anhaltspunkte und Versuche dahin gehend, die Gesichter von Spielerinnen und Spielern auf die von Spielfiguren zu übertragen. Als Beispiel hierfür gilt das Weltraumsimulationsspiel *Star Citizen*, bei dem die Gesichtsausdrücke der Spieler (allerdings ohne Rückgriff auf Deepfake-Technologien) erfasst und auf den Avatar übertragen werden.

Auch **komödiantische und satirische Verwendungen** von Deepfake-Technologien sind der Unterhaltungsdimension hinzuzurechnen. Etwa das satirische Deepfake-Video des ehemaligen brasilianischen Präsidenten Bolsonaro, worin dieser die Bevölkerung vermeintlich über die Wichtigkeit des Händewaschens während einer globalen Pandemie aufklärt, obwohl er während der Pandemie eigentlich als Gegner von Schutzmassnahmen aufgetreten war (Ajder u.a. 2019). Eines der populärsten komödiantischen Deepfake-Bilder zeigt den Papst Franziskus in weisser Daunenjacke, wie sie sonst häufig von sog. Gangster-Rappern getragen werden. Diese satirischen und komödiantischen Verwendungen können unter den Schutz der Meinungs-, Informations- und Kunstfreiheit fallen (vgl. Kapitel 4.1.1), werden allerdings angesichts ihrer Fähigkeit, Massen zu täuschen, wie es insb. im Falle des Papstes in Daunenjacke der Fall war, auch kritisiert (Keller 2023).

Die Erschaffung von **Jenseits-Charakteren** (Avatare) wird auch auf Ebene des Individuums durch Start-ups der sog. Deathtech-Branche vorangetrieben. Deepfakes von Verstorbenen sollen in der Trauerbewältigung von Nutzen sein, in dem sie als Werkzeug für die Verarbeitung von Verlusten dienen. Die dahinterliegende Geschäftsidee vereint die Produktion synthetischer Medien mit der Bestattungsbranche. So trafen beispielsweise in einer südkoreanischen TV-Show Eltern auf ihre verstorbene siebenjährige Tochter, die mittels Deepfake-Technologie inszeniert worden war. An künstlicher Kommunikation mit Jenseits-Charakteren via Chatbots, die die Sprache von Verstorbenen nachahmen, arbeiten viele Unternehmen. Sogenanntes **nostalgisches Entertainment** bietet z.B. die Anwendung *Deep Nostalgia* des Anbieters MyHeritage für Ahnenforschung und Unterhaltung. Mit der App können Gesichter von Personen auf Familienfotos animiert werden. Nicht nur Bildmaterial, sondern auch Ölgemälde und Büsten lassen sich mit dieser Software generieren, sodass die Website für persönliche Ahnenforschung und Unterhaltungszwecke kreativen Raum bietet. Um Missbrauch zu vermeiden, legt das Unternehmen den Nutzenden nahe, diese Funktion nur für eigene historische Fotos zu verwenden. Nach Aussage der Plattform verlieren die Nutzenden beim Hochladen der Bilder nicht das Urheberrecht daran und auch an dem kreierten Bewegtbild habe MyHeritage keine Urheberrechte. *Deep Nostalgia* ist ein Beispiel für eine Deepfake-Anwendung, die auf individueller Ebene positive Emotionen und Erfahrungen hervorbringen kann.

Im Kontext des Diskurses um die Digital Afterlife Industry (DAI), definiert als «any activity of production of commercial goods or services that involves online usage of digital remains» (Öhman/Floridi 2017), existieren jedoch zahlreiche offene Fragen. So etwa zu den sozialen, wirtschaftlichen und regulatorischen As-

pekten, wie eine respektvolle Haltung gegenüber den Verstorbenen seitens der technikentwickelnden Industrie aussehen kann, aber auch seitens der Nutzenden (vgl. zu diesem Thema die TA-SWISS-Studie «Tod im digitalen Zeitalter»: Strub u.a. 2024).

Im weiteren Sinne könnten auch pornografische Deepfakes zu den Chancen von Deepfake-Technologien hinzugezählt werden, sofern es sich dabei um rechtlich unproblematische (vgl. Kapitel 4.2.2.5) und erwünschte Aktivitäten handelt, wenn z.B. Partner auf solche Weise ihre sexuellen Fantasien besser ausleben können (Mirchandani 2020). Die bisherige Realität der Deepfake-Pornografie ist allerdings klar von Misogynie und der unfreiwilligen Darstellung beliebiger Frauen in entsprechenden Inhalten und der Veröffentlichung solcher Inhalte über einschlägige Netzwerke geprägt (Ajder u.a. 2019; Rini/Cohen 2022).

**Tabelle 11:** Zusammenfassung der Chancen im Anwendungsfeld Unterhaltung

| **Anwendung im Bereich Entertainment** | **Chancen** |
|---|---|
| Film- und Musikindustrie | • Digitale Wiedergänger (posthum) bei Filmproduktionen<br>• Kostengünstige Produktion von Songs ohne Musiker, Studio und Management<br>• Schutz von Minderheiten und Identitäten<br>• Effizienzsteigerung bei Produktionen<br>• Einsatz von Deepfakes bei Produktionen, sodass Personengruppen und Identitäten geschützt werden |
| Fan-Art und Gaming | • Emotionalität und positive Gefühle (Fankultur)<br>• Ausleben von Fan-Fantasien |
| Satire, Parodie | • Satirische Kritik an politischen Umständen<br>• Komödiantische Darstellung öffentlicher Personen |
| Jenseits-Charaktere | • Trauerbewältigung |
| Nostalgisches Entertainment | • Kreative Entfaltung<br>• Niederschwellige Auseinandersetzung mit Ahnenforschung |
| Sexualität | • Ausleben sexueller Fantasien |

#### 7.2.2.2. Bildung und Forschung

Im Bildungsbereich bergen Deepfake-Technologien das Potenzial, den Unterricht oder die **Bildung innovativer zu gestalten** und den Schülerinnen und Schülern so Lehrinhalte durch neue visuelle und mediale Formate näher zu bringen. Synthetische Medien können historische Persönlichkeiten (bspw. Politiker und Wissenschaftlerinnen) wieder zum Leben erwecken und so den Unterricht ansprechender und interaktiver gestalten (Westling 2019). Nach Temir (2020) können Deepfakes insgesamt das Interesse an Bildung steigern. Er sieht die Vorteile insbesondere in Fernunterrichtssystemen, die mittels synthetischer Medien interessanter gestaltet werden könnten (ebd.). Als Beispiel für eine Anwendung aus dem Bildungsbereich gilt die Gesichtsanimationssoftware *CrazyTalk* von Reallusion. Die Animation von Gesichtern aus beispielsweise Cartoons könnte von Pädagogen zur Erreichung von Lernzielen eingesetzt werden (Chesney/Citron 2019). Samsung (Neon Project) arbeitet gegenwärtig an 2D-Avataren, die dazu beitragen könnten, Tutoring- und Nachhilfesysteme zu entwickeln (Craig 2020). Diese Systeme sollen das Lernverhalten von Schülerinnen und Schülern analysieren und für die Vermittlung von Inhalten personalisierte Strategien finden. Ziel soll es sein, die Leistung der Schülerinnen und Schüler zu verbessern und ihre Motivation zu steigern. Unklar ist, inwieweit hierbei Risiken für die Privatsphäre bestehen oder welche psychologischen Auswirkungen (denkbar wäre Selbstisolation) die Technologie auf das Verhalten von Schülerinnen und Schülern haben kann. Diese Ganzkörper-Avatare könnten bald auch in anderen Bereichen als realistische virtuelle Assistenten dienen und auch in smarte Lautsprecher, wie Amazon Echo oder Google Home, integriert werden, damit für den Nutzenden das Gefühl entsteht, mit einem (vertrauten) Menschen zu sprechen (van Huijstee u.a. 2021: 16 f.).

Prominente Beispiele zeigen, dass Deepfake-Technologien als **Instrument der Medienpädagogik** für eine Bewusstseinsbildung und Sensibilität für Des- und Misinformationen verwendet werden können. So diente die Deepfake-Weihnachtsansprache der britischen Queen, als sie noch gelebt hat, dazu, die Bevölkerung auf die Gefahren desinformativer Inhalte im Netz aufmerksam zu machen. Zum einen wird die Technologie damit in der Öffentlichkeit bekannter und zum anderen werden die Rezipienten mit der kritischen Reflexion von konsumierten Informationen und Medieninhalten konfrontiert. In anderen Worten: eine Aufwertung der technischen Möglichkeiten und Wirkungen durch Umdeutung im medienpädagogischen Kontext. Ein weiteres Beispiel hierfür ist das Mark-Zuckerberg-Video, welches von einem Künstlerduo sowie der Firma Canny AI produziert worden ist. Mark Zuckerberg spricht in diesem Video nicht selbst,

sondern ein Deepfake-Zuckerberg warnt die Öffentlichkeit vor dem Missbrauch personenbezogener Daten (Beuth 2019). Auch dies gilt als Beispiel dafür, dass synthetische Medien selbst auf die «digitale Beeinflussungsindustrie» aufmerksam machen können (Kuhn 2019). Es lässt sich festhalten, dass je nach Qualität des Deepfakes die Bewusstseinsbildung und das Gelingen einer Identifizierung von falschen Informationen (Informationskompetenz) nicht vorausgesetzt werden kann. Werden Deepfakes als Instrument zur Bewusstseinsbildung eingesetzt, so sollten diese offensichtlich gekennzeichnet werden. Auch dies stellt eine Herausforderung für die Medienpädagogik dar.

Das Projekt «DeepMemory» der Universität von Amsterdam setzt Deepfakes im **Gesundheitsbereich**, der Psychologie, ein, um Angststörungen zu therapieren. Dazu wird das Gesicht des Patienten in eine Situation eingefügt, bei der dieser normalerweise Angst hat – beispielsweise das Balancieren in Höhen. Damit entstehe «eine ungesehene Perspektive des Selbst» (Ettema 2021) und der Patient könne mit den eigenen Ängsten konfrontiert werden. Für Psychologen bietet sich so der Vorteil, dass sie objektiv wahrnehmen können, was in solch einer Situation mit dem Patienten passiert. Darüber hinaus gibt es erste bekannte Fälle, in denen Personen, die beispielsweise durch Krebs ihre eigene Stimme verloren haben, mittels Deepfake-Audios ihre Stimme zurückgegeben werden konnte (Forbes 2021).

Deepfakes können auch zur Aufklärung von Verbrechen eingesetzt werden. Die Polizei in den Niederlanden setzte die Technik ein, um ein **Verbrechen aufzuklären**. Ein 2003 in Rotterdam ermordeter Junge wurde in einem Deepfake-Video auf Basis von Bildmaterial, welches der Polizei durch die Angehörigen zur Verfügung gestellt worden war, nachgestellt. Das Mordopfer wandte sich posthum an TV-Zuschauende, um Zeugen ausfindig zu machen.

**Tabelle 12:** Zusammenfassung der Chancen im Anwendungsfeld Bildung

| Anwendung im Bereich Bildung | Chancen |
|---|---|
| Innovative Bildung und Unterrichtsgestaltung | • Anregende Unterrichtsgestaltung durch Medienunterstützung<br>• Personalisierte Nachhilfe durch Avatare für bessere Lernergebnisse<br>• Steigerung der Lernmotivation |
| Deepfakes als Instrument der Medienpädagogik | • Bewusstseinsbildung und Sensibilität für Des- und Misinformationen |

#### 7.2.2.3. Werbung

Das Start-up Rephrase.AI (Indien) hat mit den Anwendungen *Rephrase Personalized* und *Rephrase Studio* Deepfake-Software auf den Markt gebracht, die es Unternehmen ermöglicht, den Trend der **Personalisierung und Individualisierung** via Deepfakes in eigene Absatzstrategien zu integrieren. Weitere Möglichkeiten bieten Deepfake-Technologien für den kommerziellen Bereich bei der Formulierung von personalisierten Werbebotschaften. Die Software *Synthesia* (www.synthesia.io) erstellt aus Texten Videos für z.B. Nachrichten oder Erinnerungen an abgebrochene Einkäufe. Deepfakes bieten hier neue kommerzielle Anwendungen in der Unternehmenskommunikation und Kundenbindung. Jedoch könnten derartige Produkte auch zur niedrigschwelligen Generierung irreführender Inhalte missbraucht werden.

Darüber hinaus kann die synthetische Produktion von Medieninhalten durch Algorithmen Unternehmen einen wesentlichen Kostenvorteil in der **Präsentation von Mode** verschaffen. Ein Beispiel hierfür ist der Stockfoto-Provider *Generated Photos*, der mittels KI-Fotos von Menschen auf Basis von bestehendem Bildmaterial generiert. Dadurch könnten Kosten für das Booking von Models reduziert werden. Weitere Potenziale ergeben sich aus virtuellen Influencern, die schon heute in den sozialen Medien wirken und denen Millionen Menschen folgen (Hiort 2022).

Nicht nur ergeben sich für Unternehmen hierbei reichweitenstarke Kooperationspartnerschaften, sondern auch in der Präsentation der Produkte effiziente Werbeproduktionen: Avatare benötigen nicht wie andere Models mehrere Takes und Shooting-Tage. Unternehmen könnten auch ihre ganz eigenen Avatare erschaffen. Denkbar ist auch, bekannte Persönlichkeiten oder Prominente virtuell erscheinen zu lassen und somit bekannte Testimonials einsetzen zu können. Hierbei sind allerdings die rechtlichen Grenzen zu beachten (vgl. Kapitel 4.2.1.1 ff.). Darüber hinaus ist es fraglich, ob mit den synthetischen Personen und Avataren *Wahrheit* als Grundsatz der Werbung angefochten werden kann.

Warner Bros setzte 2021 Deepfake-Technologie als Werbemassnahme ein, sodass interessierte Personen sich in den Trailer des Films *Reminiscence* projizieren konnten. Auf einer Website des Unternehmens konnte ein Foto hochgeladen werden, das dann durch KI-Systeme in den Trailer eingebaut wurde. Das persönliche Deepfake-Video konnte danach heruntergeladen und mit Kontakten geteilt werden. Dies verdeutlicht, wie Deepfakes als Hebel an der Schnittstelle von **Werbung und Fan-Ökonomie** wirken können, indem positive Emotionen, Identifikation mit dem Produkt und zugleich Netzwerkeffekte erzeugt werden.

Ein Beispiel dafür, dass Deepfakes auch für **öffentlichkeitswirksame Kampagnen**, die beispielsweise auf Gesundheitsförderung oder Prävention abzielen, eingesetzt werden können, ist die Kampagne «Malaria must die» (2019) mit dem ehemaligen Fussballstar David Beckham. Eine synthetische Version des Fussballers spricht darin in neun verschiedenen Sprachen, sodass das Video in vielen Ländern gezeigt werden konnte, mehr Personen erreicht und somit mehr Spenden gesammelt werden konnten (Langhart 2021).

**Tabelle 13:** Zusammenfassung der Chancen im Anwendungsfeld Werbung und kommerzielle Nutzung

| **Anwendung im Bereich Werbung und kommerzielle Nutzung** | **Chancen** |
|---|---|
| Personalisierung und Individualisierung | • Kommerzielle Anwendungen in der Unternehmenskommunikation<br>• Kundenbindung |
| Präsentation von Produkten (bspw. Mode) | • Positionierung als innovatives Unternehmen<br>• Erzielung hoher Reichweiten durch virtuelle Influencer<br>• Kosten- und zeiteffiziente Werbeproduktionen |
| Werbung und Fan-Ökonomie | • Identifikation für Fans mit Produkten/Filmen steigern<br>• Absatz steigern<br>• Netzwerkeffekte für mehr Resonanz auf Produkte/Film |
| Öffentlichkeitswirksame Kampagnen | • Internationale Kommunikation wird ermöglicht<br>• Berühmte Testimonials können auf einfachem Wege Teil der Kampagne werden und erzeugen grosse Aufmerksamkeit |

*Zwischenfazit*

Das Kapitel zeigt auf, dass Deepfakes durchaus Chancen in den Anwendungsbereichen *Unterhaltung, Bildung/Wissenschaft* und *Werbung* mit sich bringen können und viele entsprechende Anwendungen schon heute existieren. Diese Potenziale des Technologieeinsatzes sind in der gegenwärtigen, auf Risiken fokussierten Deepfakes-Debatte weitgehend unterrepräsentiert. Darüber hinaus

verstellt eine einseitige Fokussierung auf gesellschaftliche und politische Risiken den Blick auf weitere Herausforderungen, die sich aus den technologiebedingten Chancen ergeben. Denn in der Diskussion der möglichen Chancen zeigte sich, dass mit diesen häufig neue Herausforderungen einhergehen.

## 7.3. Szenarien zu Deepfakes in der Wirtschaft

Die folgenden Szenarien zu Deepfakes in der Wirtschaft lassen sich – auch wenn zu erwarten ist, dass bei vielen Szenarien ebenenübergreifende Effekte eintreten können – grundsätzlich auf drei Wirkungsebenen verteilen: Individuum, Organisation sowie Wirtschaftssystem. Die Bestimmung eines Szenarios basiert auf dem übergeordneten Effekt eines Deepfake-Einsatzes. Wenn der übergeordnete Effekt die Marktmanipulation ist, sind die Rufschädigung eines Unternehmensvorstandes oder des Unternehmens als untergeordnete Effekte zu betrachten.

Die Szenarien von Deepfakes in der Wirtschaft basieren auf denselben Kategorien wie schon die Szenarien in der Politik. Im Folgenden werden die grundlegenden Parameter der Szenarien daher in stark komprimierter Form vorgestellt (für eine erschöpfende Diskussion, vgl. Abschnitt 6.3). Im Anschluss werden die Szenarien in einer zusammenfassenden Gesamtübersicht dargestellt. Die ausführlichen Szenarien sind im Anhang zu finden (vgl. A.6). Abschliessend werden mögliche Massnahmen zum Schutz und zur Schadensbegrenzung (vgl. Kapitel 7.4) erörtert.

#### Angriffstypen und erforderlicher Ressourcenaufwand

Bei Deepfake-basierten Angriffen gegen die Wirtschaft unterscheiden wir dieselben fünf grundlegenden Angriffstypen, denen sich Deepfake-Techniken zuordnen lassen:

1. Gefälschte private Aussagen oder Handlungen
2. Social-Engineering-Angriff
3. Überwindung von Sicherheitsmassnahmen
4. Synthetische Social Botnets
5. Deepfakes von Objekten und Situationen

**Adressatenkreis und Verbreitungsweise**

Die nähere Betrachtung des Adressatenkreises dient der Bestimmung der Verbreitungsweise eines Deepfake-Inhalts. Szenarien mit einem engen Adressatenkreis bedingen in der Regel eher private Kommunikationskanäle, während Szenarien mit einem weiten Adressatenkreis darauf ausgelegt sind, mittels einer möglichst weiten Streuung möglichst viele Menschen mit dem entsprechenden Inhalt zu erreichen.

### 7.3.1. Kurzzusammenfassung der Szenarien

Wir haben auf Basis der einschlägigen Literatur und den jüngsten Entwicklungen der missbräuchlichen Verwendung von Deepfakes neun Szenarien ausgearbeitet, wie Deepfakes im Bereich der Wirtschaft für intendierte wie nicht intendierte Zwecke, die mit einem Schaden einhergehen, eingesetzt werden können (für eine ausführliche Diskussion, vgl. Anhang A.6). Darunter fallen etwa Identitätsbetrug und -diebstahl, Rufschädigung von Unternehmen oder Onlinewerbebetrug mittels synthetischer Profile. Zudem können mit Deepfakes Mitarbeitende getäuscht und biometrische Authentifizierungssysteme überwunden werden, sodass Angreifer Zugang zu gesicherten Systemen erhalten und so z.B. zu Erpressungszwecken Ransomware einspielen oder Wirtschafts- und Industriespionage betreiben können. Mit Fortschritten in Deepfake- und verwandten KI-Technologien wird schliesslich auch die Marktmanipulation und Manipulation von demokratischen Entscheidungsprozessen zulasten der Wirtschaft zunehmend möglich (vgl. auch Tabelle 14).

Auch bei den Szenarien zu Deepfakes in der Wirtschaft machen wir aufgrund der schwierigen Plausibilisierbarkeit keine Aussagen zu Eintrittswahrscheinlichkeiten. Allerdings sei erwähnt, dass einige der hier diskutierten Szenarien mit einem relativ geringen Aufwand umsetzbar sind. Hierzu zählt insbesondere das Klonen von Stimmen, um Privatpersonen oder Unternehmen zur Zahlung von Geldsummen zu manipulieren.

Im Unterschied zu den Politik-Szenarien haben die hier diskutierten Szenarien häufig einen engen Adressatenkreis. Schliesslich soll beim Identitätsbetrug, der Initiierung von Finanztransaktionen oder dem Abgreifen vertraulicher Informationen ein einzelnes Ziel in die Irre geführt werden. Dies verweist auf die Bedeutung der Entwicklung organisations- bzw. unternehmensinterner Massnahmen zum Schutz und zur Schadensbegrenzung, die von Massnahmen auf Ebene des Individuums und Gruppen bis hin zur Unternehmensspitze reichen (vgl. auch die Erörterungen im folgenden Kapitel 7.4).

**Tabelle 14:** Szenarien zu Einsatzmöglichkeiten von Deepfakes mit ökonomischem Impact (eigene Zusammenstellung)

| Ebene | Szenario | Übergeordnetes Ziel | Angriffstyp | Adressatenkreis |
|---|---|---|---|---|
| Individuum / Mitarbeitende | 1 | Identitätsbetrug | Überwindung von Sicherheitsmassnahmen | Enger Adressatenkreis |
| | | | Gefälschte Aussagen oder Handlungen | |
| | | | Social-Engineering-Angriff | |
| Organisation | 2 | Identitätsbetrug gegenüber Unternehmen | Überwindung von Sicherheitsmassnahmen | Enger Adressatenkreis |
| | | | Gefälschte Aussagen oder Handlungen | |
| | 3 | Rufschädigung eines Unternehmens mittels eines Deepfakes | Gefälschte Aussagen oder Handlungen | Eher weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | | | Vollständig synthetische Inhalte | |
| | 4 | Initiierung von Finanztransaktionen | Social-Engineering-Angriff | Enger Adressatenkreis |
| | | | Gefälschte Aussagen oder Handlungen | |
| | 5 | Deepfake-basierter Ransomware-Angriff zur Erpressung von Geld | Gefälschte Aussagen oder Handlungen | Enger Adressatenkreis |
| | | | Social-Engineering-Angriff | |
| | 6 | Abgreifen von vertraulichen Informationen (Wirtschafts- und Industriespionage) | Social-Engineering-Angriff | Enger Adressatenkreis |
| | | | Überwindung von Sicherheitsmassnahmen | |
| | | | Gefälschte Aussagen oder Handlungen | |
| | 7 | Onlinewerbebetrug mittels synthetischer Profile | Gefälschte Aussagen oder Handlungen | Weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |

| Ebene | Szenario | Übergeordnetes Ziel | Angriffstyp | Adressatenkreis |
|---|---|---|---|---|
| Markt | 8 | Marktmanipulation | Gefälschte Aussagen oder Handlungen | Weiter Adressatenkreis |
| | | | Synthetische Social Botnets | |
| | 9 | Digitales Astroturfing: Beeinflussung von demokratischen Entscheidungsprozessen zulasten der Wirtschaft | Synthetische Social Botnets | Enger bis weiter Adressatenkreis |
| | | | Gefälschte Aussagen oder Handlungen | |

## 7.4. Massnahmen zum Schutz und zur Schadensbegrenzung

Wie die vorangegangenen Szenarien verdeutlicht haben, können Deepfakes für eine Vielzahl von schädlichen Zwecken verwendet werden. So können Deepfakes nicht nur eine Bedrohung für den freien und faktenbasierten Diskurs darstellen, der die Grundlage demokratischer Gesellschaften bildet. Die Verbreitung von Deepfakes kann zudem schwerwiegende Implikationen für Politikerinnen und Politiker, für politische Institutionen wie Parteien, staatliche Stellen usw., für Unternehmen, Verbände usw. sowie die Gesellschaft insgesamt mit sich bringen. In diesem Zusammenhang stellt sich die Frage nach den Strategien und Massnahmen, die angewendet werden können, um den Herausforderungen von Deepfakes effektiv zu begegnen. Dieses Kapitel zeigt Möglichkeiten auf, wie auf Ebene des Individuums, der Organisation und der Gesellschaft auf Deepfakes reagiert werden kann, die in Schädigungsabsicht verbreitet wurden.

Die Zusammenstellung der folgenden, präventiven und reaktiven Massnahmen basiert auf der Auswertung der einschlägigen Literatur und erhebt keinen Anspruch auf Vollständigkeit. Das impliziert zweierlei: Erstens handelt es sich beim Thema Deepfakes um eines, das sich im schnellen Wandel befindet. Technologische Fortschritte und gesellschaftliche Veränderungen können die Relevanz der vorgestellten Massnahmen steigern wie reduzieren. Zweitens gibt es keine «Silberkugel» gegen Deepfakes: Denn einzelne Massnahmen (etwa Detektoren) können fehlschlagen oder unzureichende Ergebnisse liefern. Daher ist stets

eine Mischung aus unterschiedlichen Massnahmen auf unterschiedlichen Ebenen (Individuum, Organisation(en), Gesellschaft) geboten, um mögliche schädliche Auswirkungen durch Deepfakes idealerweise bereits von vornherein zu vermeiden oder im Fall der Fälle abzuwehren. Je nach Szenario können unterschiedliche Massnahmen angeraten sein. Die Zuordnung der einzelnen Massnahmen zum Schutz und zur Schadensbegrenzung im Kontext von Deepfakes in der Politik ist Tabelle 16 zu entnehmen. Eine Auflistung möglicher Massnahmen zum Schutz und zur Schadensbegrenzung im Kontext von Deepfakes in der Wirtschaft kann Tabelle 17 entnommen werden. Einige der Massnahmen, die erforderlich wären, liegen zudem nicht in der Macht der betroffenen Individuen und Organisationen, sondern betreffen andere Akteure, etwa die Politik oder Plattformbetreiber.

### 7.4.1. Sensibilisierung von Mitarbeitenden

Eine Reihe von Massnahmen, die auf der individuellen Ebene sowohl präventiv als auch reaktiv getroffen werden können, betreffen das Themenspektrum der «Sensibilisierung von Mitarbeitenden». Organisationen bzw. politische Institutionen und Parlamentarierinnen und Parlamentarier können derartige Massnahmen beispielsweise im Rahmen von Schulungen flächendeckend bekannt machen.

Im Kontext der Prävention von Deepfakes können grundlegende Sicherheitsvorkehrungen vonseiten des Individuums getroffen und ein Angriffsrisiko vermindert werden. Neben der grundlegenden Reflexion, ob Bildmaterial und Informationen veröffentlicht werden sollten (Datensparsamkeit), bedarf es einer kritischen Prüfung der Inhalte, die geteilt werden (Stimme, Umgebung, zusätzliche Information etc.). Insbesondere in den sozialen Netzwerken, wo auch Deepfakes kursieren, gilt es, das Risiko für Social-Engineering-Angriffe zu mindern. Das Bundesamt für Sicherheits- und Informationstechnik (BSI) empfiehlt, professionelle und private Social-Media-Profile getrennt zu halten. Auch sollten Freundes- und Kontaktanfragen kritisch geprüft und keine Informationen über Arbeitgeber (Ministerium, Verwaltung etc.) geteilt werden. Schliesslich sollte die ungeschützte Weitergabe persönlicher Daten in offenen, ungesicherten Netzen generell vermieden werden.

Ein weiterer elementarer Schritt zur Prävention von Deepfake-Angriffen ist die **Überprüfung der Vertrauenswürdigkeit von Absendern und Kanälen**. Das Zustandekommen des Fake-Telefonats zwischen dem Bürgermeister von Kiew und verschiedenen europäischen Amtskollegen (Schenten 2022) macht dies

deutlich. Die Kontaktaufnahme erfolgte per E-Mail, die auch in anderen Zusammenhängen ein Einfallstor für Cyberkriminelle darstellt (siehe Kapitel 7). Hier gilt es, den Absender der Mail kritisch zu prüfen. Sollte der Verdacht bestehen, dass es sich um einen Angriffsversuch handeln könnte, sollte auf die E-Mail überhaupt nicht reagiert, Anhänge nicht geöffnet und diese gelöscht werden. Im Zweifel lohnt es sich, eine Information über die Mail an die zuständige IT-Abteilung zu übermitteln. Darüber hinaus könnten aufseiten des Individuums Recherchen über den (unbekannten) Absender der Mail getätigt werden. Ist der Inhalt der Mail kryptisch und ungewöhnlich, der Absender jedoch bekannt, könnte ein Anruf bei der Absenderin getätigt werden, um sich davon zu vergewissern, dass es sich um eine seriöse Anfrage handelt (BSI 2021b). Neben der Vertrauenswürdigkeit der externen Kommunikation gilt es, auch die innerorganisationalen Kommunikationswege eindeutig und verifizierbar zu organisieren, woran alle Akteure eines Arbeitsbereichs beteiligt sind. Hierzu gehört beispielsweise, dass Verteilerlisten vorhanden sind und eine Kommunikation über dienstliche Inhalte via privater Mailadressen möglichst vermieden wird.

Angesichts der Möglichkeit von Deepfake-Anrufen gilt es, insbesondere auf die **Tonqualität und Unstimmigkeiten in der Artikulation** der Person zu achten, etwa ob die Person auf einmal eine andere Sprache spricht, mit Akzent redet (Fall Klitschko) oder sich im Allgemeinen anders als sonst artikuliert bzw. verhält. Handelt es sich um fremde Personen inner- und ausserhalb der Organisation, können keine Vergleiche angestellt werden. Daher kann es hilfreich sein, weitere Kolleginnen und Kollegen zurate zu ziehen und ggf. einen Rückruf zu tätigen. So kann zum einen die Person durch eine kurze Recherche verifiziert werden. Zum anderen ist davon auszugehen, dass Deepfake-Angreifer aufgrund der damit verbundenen Organisations- und Bereitstellungskosten eher selten eine persönliche Nummer für einen Rückruf herausgeben werden. Gerade im Falle politisch motivierter Deepfakes können Angreifer über enorme (z.B. staatliche) Ressourcen verfügen. In so einem Fall könnten die Angreifer auf etwaige Kontrollanrufe vorbereitet sein, indem sie stets ein Stimm-Double, passende Deepfake- Audios, usw. bereithalten. Schliesslich ist ein persönliches Treffen bei Unsicherheiten vorzuziehen.

In Bezug auf die Detektion von Deepfakes lassen sich drei Indikatoren feststellen, nach denen synthetische Medien erkannt werden könnten: **Kontext, audiovisuelle (technische) Schwächen und Inhalt** (Appel/Prietzel 2022). Die Kontextinformation spielt eine grosse Rolle zur Identifikation von Deepfakes. So werden Videos, die auf Deepfakes und ihre Gefahren aufmerksam machen sollen (bspw. Barack Obama und Queen Elisabeth II.), in journalistische Bei-

träge eingebettet. Fehlt diese Einbettung, so gilt es seitens der Rezipienten, die Quelle hinter den Deepfakes zu hinterfragen. Hinweise darauf, dass eine Person oder ein Objekt (Inhalt des Deepfakes) synthetisch generiert ist, geben Inkonsistenzen wie beispielsweise die Kopfhaltung oder Unregelmässigkeiten des Hintergrunds (Umgebung). In Bezug auf die technischen Schwächen der synthetischen Medien sind unter anderem das Verhalten der dargestellten Person, beispielsweise in Bezug auf Gesichtsausdrücke oder unnatürliche Lippenbewegungen, ein asynchrones Ton- und Bildverhältnis sowie der Eindruck von unsauberem Rendering des Videos (beispielsweise eine unregelmässige Oberflächenstruktur) Anzeichen für einen Deepfake (Masood u.a. 2022). Mit der zunehmenden Qualität von Deepfakes ist allerdings davon auszugehen, dass dieser Aspekt sukzessive in den Hintergrund rücken wird (Appel/Prietzel 2022).

Insgesamt ergeben sich damit folgende Kategorien für die Detektion von Deepfakes, die jedem Betrachter eine Handreichung sein können:

**Tabelle 15:** Indikatoren zur Erkennung von Deepfakes (eigene Darstellung in Anlehnung an Appel/Prietzel 2022: 2 und Masood u.a. 2022: 31)

| Indikatoren | Kategorie | Merkmale |
|---|---|---|
| Kontext | Quelle | Wer ist Absender des Deepfakes? |
| | | Wurde das Deepfake journalistisch aufgegriffen? |
| Inhalt/Objekt | Inkonsistenzen | Sichtbare Artefakte innerhalb des Bildes, wie z.B. unstimmige Kopfhaltungen und Orientierungspunkte, usw. |
| | Umgebung | Unregelmässigkeiten im Hintergrund wie Beleuchtung und andere Details (Schatten) |
| | Verhaltensanalyse | Ungewöhnliche Gesten, Lippenbewegungen, Gesichtsausdrücke |
| | Physiologie | Biologische Signale wie fehlendes Augenblinzeln oder Herzfrequenz |
| Audiovisuelle Schwächen | Synchronisation | Zeitliche Konsistenz, z.B. Inkonsistenzen zwischen angrenzenden Bildern/Modalitäten |
| | Kohärenz | Fehlender optischer Fluss und Artefakte wie Flackern und Zittern zwischen Bildern |
| | Struktur | Rendering des Videos (bspw. eine unregelmässige Oberflächenstruktur) |

Im Massachusetts Institute of Technology (MIT) wurde die Hypothese aufgestellt, dass die Analyse des Aussehens der Deepfake-Avatare und eine Routine im Erkennen von manipulierten Medien dazu führen werde, dass in Zukunft eine Vielzahl von manipulierten visuellen Medien erkannt werden (Groh n.n.). Hierfür hostet das MIT Media Lab die Website *Detectfakes*, auf der Menschen sich mittels zahlreicher Beispiele selbst in Bezug auf die Erkennung von Deepfakes testen können (MIT Media Lab n.n.). Die Forschung zeigt, dass insbesondere Menschen, die über ein hohes Mass an analytischem Denkvermögen verfügen, überdurchschnittlich gut in der Erkennung von Deepfakes sind (Appel/Prietzel 2022). Andere Forschungsarbeiten nähren hingegen berechtigte Zweifel an der Fähigkeit von Mediennutzenden, Deepfakes zu erkennen. So zeigen Köbis u.a. (2021), dass die eigene Fähigkeit, Deepfakes zu erkennen, überschätzt wird. Nightingale/Farid (2022) zeigen, dass Menschen synthetisch erstellte Gesichter nicht von echten unterscheiden können und ersteren sogar mehr vertrauen. Auch unsere eigenen Studienergebnisse zeigen, dass Menschen häufig nicht in der Lage sind, gut bis sehr gut gemachte Deepfakes zu erkennen. Wichtig wäre also, übergeordnete Medienkompetenzen zu vermitteln und das generelle kritische Denken anzuregen (vgl. Kapitel 3).

Konkrete Schritte, die Individuen befolgen können, sind folgende:

1. **Erste Anzeichen eines irreführenden Inhalts**: Löst ein Video oder eine Abbildung beim Betrachten Emotionen wie Wut oder Empörung aus? Und/oder handelt es sich dabei um einen sensationserheischenden Inhalt? Könnte mit dem Inhalt Stimmung für eine bestimmte politische oder weltanschauliche Strömung gemacht werden? Wie sehr vertraue ich der Quelle? Wenn ich der Quelle vertraue: Passt der Inhalt zu anderen Meldungen der Quelle oder könnte die Quelle kompromittiert worden sein?
    - Die Beantwortung einer dieser Fragen mit Ja kann ein Anstoss sein, den Wahrheitsgehalt des entsprechenden Inhalts näher zu untersuchen. Klar ist aber auch: Schon derartige Tipps werden voraussichtlich das Problem der Erosion der Glaubwürdigkeit realer Inhalte (truth decay) nähren. Daher sollten sie stets bloss Teil umfassenderer Aufklärungskampagnen sein.
2. Weil Deepfake-Generatoren, aber auch Text-zu-Bild-Generatoren, aktuell nicht immer realistische Ergebnisse liefern, kann auch **ein näherer, kritischer Blick auf einen jeweiligen Inhalt** erste konkretere Hinweise darauf liefern, dass es sich um eine Fälschung handelt: Einige Bildgeneratoren sind etwa nicht in der Lage, menschliche Gesichtszüge ohne unnatürliche Zuckun-

gen darzustellen, ein natürliches Augenblinzeln darzustellen, menschliche Finger akkurat abzubilden, in anderen Fällen sind Gesichter verschwommen oder Gliedmassen abgetrennt. Generell gilt: Sobald ein Bildinhalt Elemente enthält, die unnatürlich wirken, sollte Skepsis angebracht sein.

- Allerdings gilt hierbei: Auch wenn dieser Tipp kurzfristig hilfreich sein mag – auf mittel- bis langfristige Sicht werden die Werkzeuge zur Generierung gefälschter Inhalte so fortgeschritten sein, dass sie mit dem blossen menschlichen Auge nicht wahrgenommen werden können.

Wenn Zweifel am Wahrheitsgehalt eines Inhalts bestehen, bieten sich mehrere weitere Möglichkeiten:

3. **Onlinedetektoren** bieten eine einfache Möglichkeit, den Wahrheitsgehalt von Bildern oder Videos zu überprüfen (vgl. auch Kapitel 2.7). Im einfachsten Fall muss lediglich der Link zu dem zu überprüfenden Video in das entsprechende Feld eingegeben werden. Auf ähnliche Weise können auch Bilder überprüft werden, etwa mit dem *AI Image Detector* von Hugging Face[288] (https://huggingface.co/spaces/umm-maybe/AI-image-detector).
   - Generell gilt bei den Detektoren, dass sie mit Wahrscheinlichkeitswerten arbeiten und deshalb regelmässig keine abschliessende Auskunft über den Wahrheitsgehalt eines Bildes oder Videos geben können. Dennoch sollten Nutzerinnen und Nutzer bereits bei einer Wahrscheinlichkeit von 50 % den Wahrheitsgehalt eines Inhalts anzweifeln.
4. **Rückwärtssuche**: Spezialisierte Anbieter (z.B. TinEye) und Suchmaschinenbetreiber (Google oder Yandex) bieten an, Bilder per sog. Rückwärtssuche daraufhin zu untersuchen, ob und wo sie im Internet bereits andernorts hochgeladen wurden.

### 7.4.2. Strukturelle Massnahmen in Organisationen

Im politischen Kontext sind für die Prävention von Deepfake-Angriffen politische Entscheidungstragende und die jeweiligen organisationalen Strukturen von Departementen und weiteren Verwaltungen relevant. Ein erster Schritt kann es sein, die Sichtbarkeit und Relevanz des Themas in den eigenen organisationalen

[288] Diese Nennungen seien als Hilfestellung verstanden. Weil der Markt für Detektoren sich noch im Entstehen befindet, ist mit Veränderungen nach der Studienpublikation zu rechnen, sodass wir empfehlen, selbstständig nach entsprechenden Anwendungen zu suchen.

Strukturen zu erhöhen **(Risikowahrnehmung)**. Es gilt also, die «Awareness» durch gezielte Kommunikation und Diskussion über schädliche Einsatzmöglichkeiten von Deepfakes zu stärken. Neben dem Aufbau eines entsprechenden Problem- und Sicherheitsbewusstseins der Mitarbeitenden gilt es, einen **Katalog von relevanten Gefährdungen** zu benennen und die Erwartungshaltung hinsichtlich der Informationssicherheit klar zu kommunizieren. Die im Rahmen dieser Studie erarbeiteten Szenarien können einen Anknüpfungspunkt zur Identifizierung von Gefährdungsszenarien bieten. Nicht jedes Szenario wird aber jede Organisation, Person oder Gruppe gleichermassen betreffen. Deshalb sollte die Abschätzung des Gefährdungspotenzials im Einzelfall durch die entsprechende(n) Person(en), Gruppen oder Organisationen durchgeführt werden.

Die Prävention vor Deepfake-Angriffen setzt bei der Sensibilisierung von Politikerinnen und Politikern sowie den Mitarbeitenden in Verwaltung und Ministerien an und zielt auch auf die Erhöhung der Resilienz von IT-Infrastrukturen ab (Rogge 2018). Diese Sensibilisierung kann jedoch nicht Sache der Individuen sein. Stattdessen muss die jeweilige Organisation den Aufbau individueller **Resilienz mit gezielten Programmen, Dialogformaten und Schulungen fördern**, sodass die Mitarbeitenden befähigt werden, einen Deepfake entweder zu erkennen oder skeptischer gegenüber möglicherweise manipulierten oder gefälschten Inhalten zu sein. Die **Organisationskultur** kann dazu beitragen, dass hierarchische Machtgefälle mittels Deepfakes nicht ausgenutzt werden und Mitarbeitende Zweifel über die Echtheit eines Telefonats, Video-Calls usw. auch äussern (Europäische Kommission 2021b; Espeloer 2022).

#### 7.4.2.1. Die Schaffung von oder Schulung bestehender Teams

Neben der Risikowahrnehmung ist es zudem notwendig, ein **organisationsweites Schutzkonzept** aufzubauen. Hierfür müsste nach vorangegangener Risikowahrnehmung der handelnden Personen auch die **Risikoanalyse** für den jeweiligen Arbeitsbereich erfolgen (Daloz 2021).

Für den Fall, dass politische Handlungstragende selbst Opfer eines Deepfake-Angriffs werden, bedarf es eines **Notfall- und Krisenplans**, der die interne Handlungsfähigkeit gewährleistet und externe Kommunikation miteinschliesst. Zur Realisierung einer zeitnahen Reaktion wären dann etwa entsprechende Kommunikationskanäle zu Medienorganisationen bereitzuhalten bzw. aufzubauen, um eine Klarstellung über Multiplikatoren schnell in die Gesellschaft tragen zu können. Selbst im Falle eines Ausfalls der technischen Bedingungen vor Ort (wenn zusätzlich zur Zirkulation eines Deepfakes ein Hackerangriff interne

Systeme lahmlegt, um allfällige Reaktionen zu erschweren) müsste ein solches Team immer noch reaktionsfähig bleiben. Immerhin besteht mit Smartphones die Möglichkeit, derartige Massnahmen zügig zu ergreifen, indem beispielsweise ein Video zur Richtigstellung von Falschbehauptungen an jedem beliebigen Ort aufgezeichnet und an entsprechende Multiplikatoren, insbesondere Medien, weitergeleitet werden kann.

Ein weiterer Ansatz zur Prävention von Deepfake-Angriffen im politischen Kontext ist eine klare **Verteilung von Rollen** im Arbeitsumfeld. Diese können aufgrund von technischen oder organisatorischen Fähigkeiten definiert und verteilt werden. Hausinterne **«Faktencheck-Teams»** können gegründet bzw. ausgebaut und dahin gehend eingesetzt werden, redaktionelle und recherchierende Aufgaben beispielsweise in Bezug auf Quellen und Akteure zu übernehmen.[289] Darüber hinaus können die Faktenchecker den Onlineraum kontinuierlich auf potenziell irreführende Inhalte überprüfen und eine Analyse des Desinformationsökosystems vornehmen, um ggf. das für den Notfall- und Krisenplan zuständigen Team zur Initiierung weiterer Schritte zu informieren (Juneja/Mitra 2022). Diese Teams könnten sich auch der Frage annehmen, ob und inwiefern Werkzeuge, wie die Authentifizierung von Inhalten oder die Nutzung von Deepfake-Detektoren, im Falle der eigenen Organisation ratsam wären (vgl. die Kapitel 2.6.1.1 und 2.6.2).

Ähnlich wie bei der Sensibilisierung von Mitarbeitenden bedarf es auch bei der Schaffung bzw. Schulung von Teams einer **organisationalen Entscheidung für die Realisierung der Massnahme**.

#### 7.4.2.2. Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen

Dass eine Richtigstellung seitens Betroffener aufgezeichnet und über soziale Medien unmittelbar geteilt und an Multiplikatoren weitergereicht wird, bedeutet noch nicht, dass sie auch diejenigen Menschen erreicht, die bereits durch ein Deepfake getäuscht wurden. Eine wichtige Bedingung für den Erfolg einer Richtigstellung ist, dass insbesondere Medien als Multiplikatoren wirken und entsprechende Richtigstellungen zeitnah an die Bevölkerung kommunizieren. Auch andere Akteure wie Plattformbetreiber können hierbei behilflich sein, indem die

[289] Faktenchecks werden im Verwaltungsbereich bereits im Eidgenössischen Finanzdepartement (EFD) vorgenommen und veröffentlicht (EFD, 2023). Auch im Journalismus sind Faktenchecks ein etabliertes Mittel (vgl. Kapitel 5.3.2).

Reichweite derartiger Richtigstellungen erhöht wird. Für eine noch grössere Reichweite bedürfte es zudem der Unterstützung aus der (Zivil-)Gesellschaft, indem engagierte Bürgerinnen und Bürger einen schädigenden Deepfake-Inhalt z.B. direkt in Chatgruppen oder Social-Media-Postings mit der Realität bzw. der Richtigstellung konfrontieren.

Während das Problem der Reichweite bzw. Erreichung der Öffentlichkeit also durch zügiges Reagieren von Multiplikatoren adressiert werden kann, stellen sich aufseiten der Rezipienten grundlegendere Herausforderungen: Denn dass die Richtigstellung viele Menschen erreicht (insb. solche, die bereits zuvor das Deepfake gesehen haben), bedeutet nicht, dass der Richtigstellung auch geglaubt wird. Es bedeutet auch nicht, dass das Deepfake nicht trotzdem einen bleibenden Eindruck hinterlässt. In einer Studie konnte aufgezeigt werden, dass Informationen von Plattformen, die auf ein Deepfake hinweisen, nicht dazu führten, dass die Studienteilnehmenden manipulierte Videoinhalte besser erkennen konnten. Die Wissenschaftler warnen, dass solche Hinweise und Warnungen gar den *Backfire-Effekt* auslösen, also zu mehr Misstrauen in Medien im Allgemeinen führen (Ternovski u.a. 2021: 8 ff.). Unsere Studienergebnisse bestätigen zwar, dass Hilfestellungen zum Erkennen von Deepfakes wirkungslos sind, allerdings konnten wir auch keinen Backfire-Effekt feststellen (vgl. Kapitel 3).

#### 7.4.2.3. Das Treffen von grundlegenden IT-Sicherheitsvorkehrungen

IT-Sicherheitsvorkehrungen kommt in einer Zeit ständig zunehmender Hackerangriffe und IT-Sicherheitsvorfälle eine grundsätzlich wachsende Bedeutung zu. Speziell im Hinblick auf den Schutz vor Deepfakes kommen IT-Sicherheitsvorkehrungen vor allem im Hinblick auf die Abwehr von Deepfake-Angriffen, welche die Überwindung biometrischer Sicherheitsmassnahmen zum Ziel haben, sowie bei Social-Engineering-Angriffen infrage. Moderne Zugriffsrechtemanagementsysteme z.B. können den Zugang zu (sensiblen) Daten und Diensten im Falle eines Social-Engineering-Angriffs reduzieren.

Hinsichtlich der Massnahmen zur Abwehr von Cybergefahren existiert ein reichhaltiger und hilfreicher Fundus an Fachliteratur. An dieser Stelle sei lediglich beispielhaft auf entsprechende Hilfestellungen von behördlichen Stellen wie der Agentur der Europäischen Union für Cybersicherheit (ENISA), der US-amerikanischen Cybersicherheitsbehörde CISA und Unternehmen wie Symantec oder SwissInfosec hingewiesen.

Organisationen, die zur Autorisierung auf Gesichts- oder Sprachbiometrie setzen, ist zu empfehlen, den gesamten Autorisierungsprozess zu überarbeiten.

Europol empfiehlt etwa, anstelle von Deepfake-Detektoren besser die Robustheit der bestehenden Systeme zu verbessern. Darunter:

- Nicht allein auf Audio-Authentifizierung setzen, sondern auf audiovisuelle Authentifizierung.
- Verwendung einer Live-Videoverbindung.
- Durchführung von zufälligen komplizierten Handlungen in Echtzeit vor der Kamera, etwa das Bewegen der Hand vor das Gesicht.

Zur Erkennung des Versuchs eines Identitätsbetrugs mittels Gesichtsbiometrie-Spoofing sollten die Systeme auf fortschrittlichste Authentifizierungsmassnahmen setzen (Wojewidka 2020: 6 f.), darunter:

- Erkennen können, ob das im Aufnahmefeld befindliche Objekt 3D oder 2D ist, also ob es sich um einen menschlichen Kopf handelt oder um ein Blatt Papier mit einem Gesichtsfoto.
- Das System sollte mittels *liveness detection* (die von verlässlichen Third-Party-Anbietern zertifiziert wurde) erkennen können, ob das im Aufnahmefeld befindliche 3D-Objekt lebendig ist. Auf diese Weise könnte vermieden werden, dass 3D-Gesichtsmasken fälschlicherweise als authentisches Gesicht erkannt werden.
- Sollte das Live-3D-Video mit früheren Aufnahmen einer 3D-Facemap abgeglichen werden, das eine akkurate digitale Abbildung eines 3D-Fotos eines menschlichen Gesichts darstellt.

#### 7.4.2.4. Freiwillige Meldung von Deepfake-Vorfällen durch betroffene Organisationen

Nach Bekanntwerden des ersten Deepfake-Voice-Phishing-Falls wurde vielfach über dieses Thema berichtet und das Treffen von geeigneten Schutzmassnahmen zur Verhinderung vergleichbarer Angriffe bei anderen Unternehmen war eins der prioritär behandelten Aspekte. Die Schweizer Strategie zum Umgang mit Cybergefahren ist dezentral und verteilt sich je nach Aufgabengebiet auf unterschiedliche Schultern. Die Verantwortung für den Eigenschutz verbleibt bei den Unternehmen selbst. Der Bereich der Cyberstrafverfolgung, was also insb. die Meldung und strafrechtliche Verfolgung von Zwischenfällen betrifft, liegt primär bei den Kantonen. Dabei nimmt das BACS eine beratende Rolle ein. Je nach Fall können auch das Bundesamt für Polizei (fedpol) und die Bundes-

anwaltschaft (BA) zuständig sein (Bundesrat 2023b: 9 ff.). Nebst der für 2025 geplanten Einführung einer Meldepflicht bei Cyberangriffen auf kritische Infrastrukturen kann auch das freiwillige Melden von Deepfake-basierten Vorfällen von Vorteil sein.

Das Melden eines Angriffs kann zunächst dem betroffenen Unternehmen bei der Schadensbegrenzung helfen, etwa indem die zuständigen Behörden Hilfestellungen geben. Das Bekanntwerden von Angriffsvektoren kann ausserdem bei anderen Organisationen zu erhöhter Aufmerksamkeit, der Verbesserung von Schutzmassnahmen und damit einer verringerten Chance des wiederholten Erfolgs desselben Angriffstyps führen. Als Meldestelle kommen neben den zuständigen Polizeien auch das BACS oder private Beratungsagenturen infrage. Bislang tauchen Deepfakes in den Berichten des BACS nicht als eigene Kategorie auf. Sofern Deepfake-basierte Angriffe gemeldet bzw. von den kantonalen Dienststellen an das BACS weitergegeben werden, könnte die klare Benennung dieser als solche Aufmerksamkeit aufseiten der Unternehmen für die von Deepfakes ausgehenden Gefahren schaffen.

#### 7.4.2.5. Fortlaufende Massnahmen zur Betrugsprävention

Die Täuschung von IT-Systemen oder von Menschen durch Deepfakes oder die Beeinflussung von demokratischen Entscheidungsprozessen erfordert präventive als auch reaktive Abwehrmassnahmen.

Bei wichtigen politischen Entscheidungsprozessen, die von öffentlichen Konsultationsprozessen begleitet werden, können Verbände, zivilgesellschaftliche Akteure usw. eine wichtige Rolle einnehmen, indem sie die Überprüfung der Bevölkerungseingaben (sofern diese öffentlich zugänglich sind) nicht allein den politischen Verantwortlichen überlassen, sondern diese ihrerseits auf ihre Authentizität hin überprüfen. Bei – mittel- bis langfristig zu erwartenden – ausreichend ausgefeilten Eingaben dürfte dies jedoch wenig Erfolg versprechend sein. Im Zweifelsfall könnten eigene repräsentative Bevölkerungsumfragen in Auftrag gegeben werden, um einem möglicherweise manipulierten Meinungsbild verlässlichere Informationen entgegenzusetzen.

#### 7.4.2.6. Weitere mögliche Massnahmen

Weitere präventive und reaktive Massnahmen, die in der Literatur diskutiert werden, liegen ausserhalb des unmittelbaren Verantwortungsbereichs jener Akteu-

re, die in Politik und Wirtschaft selbst von Deepfakes mit Schädigungsabsicht adressiert werden könnten, und bedürfen weiterer Schritte, weshalb sie im Folgenden lediglich kurz andiskutiert werden. Diese Massnahmen umfassen:

- Moderation von Inhalten durch Plattformbetreibende
- Transparenz- und Kennzeichnungspflicht
- Gesellschaftliche Sensibilisierung und Resilienzaufbau
- Internationale Normen und Regeln

Da Social-Media-Plattformen eine entscheidende Rolle bei der Verbreitung von Deepfakes spielen, sieht eine Reihe von Vorschlägen Massnahmen auf Ebene der Plattformregulierung in Form der **Moderation von Inhalten durch Plattformbetreibende** vor, um die Rechte von Betroffenen besser zu schützen, die Verbreitung von Deepfakes einzudämmen und zugleich ungewollte Auswirkungen auf die Meinungsäusserungsfreiheit zu vermeiden (vgl. hierzu die Diskussion von Regulierungsansätzen in Kapitel 4.5.2).

**Transparenz- und Kennzeichnungspflichten** werden ebenfalls mit Blick auf Plattformregulierung diskutiert (vgl. Kapitel 4.5.2). Die Diskussion von Authentifizierungsverfahren in Kapitel 2.6.1.1 zeigt zudem, dass zwar derzeit noch keine verlässlichen Verfahren bestehen, aber auch seitens der Informationswirtschaft intensiv an weiteren Lösungen geforscht wird.

Auf die Bedeutung der **gesellschaftlichen Sensibilisierung und des Resilienzaufbaus** verweisen sowohl Literatur als auch unsere empirischen Studienergebnisse. Hierbei zeigte sich insb. die Bedeutung von Social-Media-Literacy (vgl. Kapitel 3). Solche übergeordneten Medienkompetenzen können das kritische Denken der Menschen aktivieren und sie für einen kritischen Umgang auch mit Deepfakes sensibilisieren, ohne dass sie überkritisch werden und auch reale Videos für eine Fälschung halten. Aus demokratietheoretischer Sicht argumentieren Zimmermann und Kohring (2018), dass Desinformation als «Symptom eines strukturellen Problems betrachtet» werden muss, das seine Wurzel im Misstrauen und der Formierung einer Gegenöffentlichkeit gegenüber demokratischen Institutionen hat (Zimmermann/Kohring 2018). Insofern kann der Aufbau von Vertrauen in demokratische Institutionen dazu beitragen, dass die Gesellschaft resilienter wird. Taiwan, das regelmässig Ziel von chinesischen Desinformationskampagnen wird, setzt u.a. auf den verstärkten Einbezug der Zivilgesellschaft bei der Bekämpfung von Desinformation. Dadurch, dass die Bevölkerung selbst in die Pflicht genommen wird, könne dem Framing, wonach der Staat über

richtig und falsch entscheide, entgegengewirkt werden, während gleichzeitig die Gesellschaft sensibilisiert und resilienter wird (Smith 2020).

Angesichts der Möglichkeit, dass Deepfakes mit Schädigungsabsicht in der Schweiz durch Akteure verbreitet werden, die selbst im Ausland beheimatet sind, stellt sich die Frage der Adressierung solcher Bedrohungsakteure. Weil der rechtliche Zugriff auf solche Akteure in einigen Fällen nur begrenzt und in anderen Fällen gar nicht möglich ist (vgl. Kapitel 4.5.1.4), ist die **Förderung internationaler Normen und Regeln** unabdingbar, sodass Regierungen unterschiedlicher Staaten sich gegenseitig die Einhaltung bestimmter Normen zusichern. Hier sind also diplomatische Massnahmen zu begrüssen, wie der AI Safety Summit Ende 2023, auf dem Vertreter von 28 Staaten, internationalen Organisationen sowie Akteure aus Wirtschaft und Zivilgesellschaft für die Gewährleistung einer friedvollen und regelbasierten Nutzung von KI-Technologien zusammengekommen sind.

**Tabelle 16:** Zuordnung von Massnahmen zum Schutz und zur Schadensbegrenzung zu den Szenarien im Bereich Deepfakes in der Politik

| Ebene | Szenario | Übergeordneter Effekt | Massnahmen |
|---|---|---|---|
| Individuum | 1 | Erpressung bzw. Einschüchterung eines Politikers | Grundlegende Sicherheitsvorkehrungen |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 2 | Rufschädigung eines Politikers | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 3 | Anstachelung zu Gewalttaten gegen einzelne Politikerinnen und Politiker | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |

| Ebene | Szenario | Übergeordneter Effekt | Massnahmen |
|---|---|---|---|
| Institutionen | 4 | Rufschädigung einer Partei oder (staatlichen) Institution | Sensibilisierung von Mitarbeitenden und Politikern |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 5 | Erbeutung von vertraulichen Informationen | Sensibilisierung von Mitarbeitenden |
| | | | Grundlegende IT-Sicherheitsvorkehrungen (einschl. sichere Authentifizierungs-verfahren) |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| Gesellschaft | 6 | Beeinflussung einer Wahl | Transparenz- und Kennzeichnungspflicht |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 7 | Beeinflussung von politischen Entscheidungsprozessen | Sensibilisierung von Mitarbeitenden und Politikern |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |

| Ebene | Szenario | Übergeordneter Effekt | Massnahmen |
|---|---|---|---|
| Gesellschaft | 8 | Verschärfung sozialer Spannungen | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 9 | Beschädigung der Demokratie | Sensibilisierung von Mitarbeitenden und Politikern |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 10 | Gefährdung der öffentlichen Sicherheit | Sensibilisierung von Mitarbeitenden und Politikern |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 11 | Beeinflussung der internationalen Beziehungen | Sensibilisierung von Mitarbeitenden und Politikern |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Gesellschaftliche Sensibilisierung und Resilienzaufbau |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |

| Ebene | Szenario | Übergeordnetes Ziel | Massnahmen |
|---|---|---|---|
| Individuum / Mitarbeitende | 1 | Identitätsbetrug | Gesellschaftliche Sensibilisierung und Resilienz |
| | | | Grundlegende IT-Sicherheitsvorkehrungen (einschl. sichere Authentifizierungsverfahren) |
| | | | Sensibilisierung von Mitarbeitenden |
| Organisation | 2 | Identitätsbetrug gegenüber Unternehmen | Fortlaufende Aktualisierung von internen Betrugspräventionsmassnahmen |
| | | | Grundlegende IT-Sicherheitsvorkehrungen (einschl. sichere Authentifizierungsverfahren) |
| | | | Sensibilisierung von Mitarbeitenden |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 3 | Rufschädigung eines Unternehmens mittels eines Deepfakes | Sensibilisierung von Mitarbeitenden |
| | | | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Gesellschaftliche Sensibilisierung und Resilienz |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Transparenz- und Kennzeichnungspflicht |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | 4 | Initiierung von Finanztransaktionen | Sensibilisierung von Mitarbeitenden |
| | | | Grundlegende IT-Sicherheitsvorkehrungen (einschl. sichere Authentifizierungsverfahren) |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 5 | Deepfake-basierter Ransomware-Angriff zur Erpressung von Geld | Sensibilisierung von Mitarbeitenden |
| | | | Grundlegende IT-Sicherheitsvorkehrungen (einschl. sichere Authentifizierungsverfahren) |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |

| Ebene | Szenario | Übergeordnetes Ziel | Massnahmen |
|---|---|---|---|
| Organisation | 6 | Abgreifen von vertraulichen Informationen (Wirtschafts- und Industriespionage) | Sensibilisierung von Mitarbeitenden |
| | | | Grundlegende IT-Sicherheitsvorkehrungen (einschl. sichere Authentifizierungsverfahren) |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | 7 | Onlinewerbebetrug mittels synthetischer Profile | Transparenz- und Kennzeichnungspflicht |
| | | | Sensibilisierung von Mitarbeitenden |
| Markt | 8 | Marktmanipulation | Schaffung von Teams, die schnelle Gegenreaktionen durchführen können |
| | | | Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | 9 | Digitales Astroturfing: Beeinflussung von demokratischen Entscheidungsprozessen zulasten der Wirtschaft | Fortlaufende Aktualisierung von internen Betrugspräventionsmassnahmen |
| | | | Moderation von Inhalten durch Plattformbetreiber |
| | | | Sensibilisierung von Mitarbeitenden |
| | | | Internationale Normen und Regeln (Verhaltenskodex) |

## 7.5. Zwischenfazit

Zunehmende Fälle der Verwendung von Deepfakes für wirtschaftskriminelle Zwecke waren Anlass des Themas Deepfakes in der Wirtschaft. Dazu wurde zunächst nach möglichen Herausforderungen und Chancen von Deepfakes in der Wirtschaft gefragt und im Anschluss nach Handlungsmöglichkeiten im Bereich von Massnahmen zum Schutz und zur Schadensbegrenzung.

### Herausforderungen von Deepfakes

Der Medien- und Literaturüberblick zu den Herausforderungen zeigte, dass die Verwendung von Deepfakes für wirtschaftskriminelle Zwecke bereits heute erfolgt. Ausgehend von den (technologischen) Einsatzmöglichkeiten von Deepfakes und der Operationsweise krimineller Gruppierungen gehen wir davon aus, dass Deepfakes auch weiterhin Eingang in das Angriffsrepertoire professioneller krimineller Akteure finden werden.

### Chancen von Deepfakes

Auf Basis einer Medieninhaltsanalyse zeigte sich ausserdem, dass der Einsatz von Deepfakes in verschiedenen Anwendungsbereichen mit positiven Nutzungszwecken verbunden ist, die in der gegenwärtigen, auf Risiken fokussierten, Deepfakes-Debatte weitgehend unberücksichtigt bleiben. Chancen von Deepfakes ergeben sich auf Basis dieser Analyse vor allem in den Bereichen Unterhaltung, Bildung sowie Werbung und kommerzielle Nutzung. Der Rückgriff auf Deepfake-Technologien in diesen Bereichen kann neue Unterhaltungsangebote schaffen, Fankultur unterstützen und damit auch die Kundenbindung stärken, Kosten- und Zeiteinsparungen bewirken, neue satirische Unterhaltungsangebote schaffen, Schutz von Minderheiten und Identitäten ermöglichen, zur Verbesserung von Bildungsangeboten genutzt werden und zur Steigerung der Lernmotivation führen oder zur Stärkung der Medienkompetenz beitragen.

### Szenarien zu Deepfakes in der Wirtschaft

Zur Konkretisierung der möglichen Nutzung von Deepfakes in der Wirtschaft wurden auch für diesen Bereich Szenarien ausgearbeitet. Diese zeigen Einsatzmöglichkeiten von Deepfakes im Bereich der Wirtschaftskriminalität, insbesondere für Identitätsbetrug und -diebstahl, Rufschädigung von Unternehmen, Onlinewerbebetrug mittels synthetischer Profile bis hin zur Marktmanipulation und Manipulation von demokratischen Entscheidungsprozessen zum Nachteil der Wirtschaft oder von bestimmten Wirtschaftssektoren. Zudem können mit Deepfakes Mitarbeitende getäuscht und biometrische Authentifizierungssysteme überwunden werden, sodass Angreifer Zugang zu gesicherten Systemen erhalten und so z.B. zu Erpressungszwecken Ransomware einspielen oder Wirtschafts- und Industriespionage betreiben können. Auch die Szenarien aus diesem Kapitel können Unternehmen, Verbände etc. bei der Erstellung eines eigenen Risikoprofils unterstützen.

### Mögliche Massnahmen zum Schutz und zur Schadensbegrenzung

Die Erörterung der Szenarien aus den Kapiteln Politik und Wirtschaft verdeutlicht, dass neben den Gefahren auch eine Reihe von Massnahmen zum Schutz und zur Schadensbegrenzung existiert, die schon mit den heute verfügbaren Mitteln bei Akteuren aus Politik und Wirtschaft gleichermassen umgesetzt werden können. Dazu zählen insb.:

- **Die Sensibilisierung von Mitarbeitenden**, um verdächtige Situationen zu erkennen und ggf. geeignete Schritte zu ergreifen, wie die Meldung eines Anrufs, bei dem ein Verdacht auf eine Deepfake-Stimme vorliegt.
- **Die Schaffung von (oder Schulung bestehender) Teams**, die in der Lage sind, schnelle Gegenreaktionen durchzuführen, etwa das Filmen und Veröffentlichen einer klarstellenden Botschaft.
- **Schnelle Verbreitung von Richtigstellungen durch Medienorganisationen**, damit die von den durch ein Deepfake Betroffenen erstellten Richtigstellungen auch schnell eine möglichst grosse Anzahl an Menschen erreichen.
- **Das Treffen von grundlegenden IT-Sicherheitsvorkehrungen** ist auch zur Abwehr schädlicher Deepfake-Einsätze geeignet, indem etwa Authentifizierungsverfahren gesichert werden und Identitätsbetrug verhindert wird.

# 8. Empfehlungen

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

Dieses Kapitel soll Gestaltungs- und Handlungsmöglichkeiten für einen nachhaltigen und verantwortungsbewussten gesellschaftlichen Umgang mit Deepfakes aufzeigen. Die Analyse der technischen, juristischen, pädagogischen, journalistischen, wirtschaftlichen und politischen Fragen hat deutlich gemacht, dass Handlungsbedarf auf unterschiedlichen Ebenen besteht.

Die Empfehlungen wurden auf Basis der Ergebnisse der einzelnen Kapitel sowie aus der Untersuchung von bereits praktizierten Regulierungsansätzen und in der Literatur geäusserten Regulierungsvorschlägen entwickelt. Dabei wurde Wert auf die praktische Realisierbarkeit der Empfehlungen im spezifischen regulatorischen Kontext der Schweiz gelegt.

Bereits in den vorangehenden Kapiteln trat teilweise zutage, dass sich die Problematik der «schädlichen» Deepfakes überschneidet mit anderen Diskussionen im Kontext digitaler Kommunikation. So werden einige der hier identifizierten Herausforderungen auch im Zusammenhang mit Desinformation diskutiert, andere sind relevant im Umfeld von Cyberkriminalität und Cyberangriffen sowie digitaler Gewalt, Mobbing und Stalking. Einige der hier präsentierten Empfehlungen adressieren deshalb auch breitere, nicht allein bei Deepfakes auftretende Problematiken.

Die Empfehlungen sind nach Adressaten gegliedert. So wird unterschieden zwischen dem *Staat als Regulierungsakteur* (Kapitel 8.1), *Gesellschaft und Bildungseinrichtungen* (Kapitel 8.2), *Organisationen in allen Branchen* (Kapitel 8.3), der *Kommunikationsbranche* (Kapitel 8.4), *Plattformbetreibern* (Kapitel 8.5) sowie *Medienorganisationen, der Medienausbildung und Nachrichtenagenturen* (Kapitel 8.6).

## 8.1. Staat als Regulierungsakteur

### 8.1.1. Plattformregulierung

Ein Grossteil der schädlichen Deepfakes wird über grosse Onlineplattformen verbreitet. Es empfiehlt sich deshalb, die bestehenden und geplanten Gesetzgebungsmassnahmen zur Regulierung von Plattformen weiterzuverfolgen. Um

sicherzustellen, dass Plattformanbieter ihre Verantwortung bei der Verbreitung von Inhalten in der Öffentlichkeit wahrnehmen, braucht es eine Regulierung, welche gleichzeitig die Meinungsäusserungsfreiheit garantiert und die Rechte der Geschädigten wahrt. Gleichzeitig ist eine Verbesserung der Rechtsdurchsetzung im internationalen Verhältnis notwendig. Dazu müssten zusätzliche Kooperationsinstrumente mit anderen Staaten geschaffen werden, und es bräuchte eine Verpflichtung der Plattformen, mit den hiesigen Strafverfolgungsbehörden bei der Aufklärung von Straftaten zusammenzuarbeiten. Während die EU mit dem Digital Services Act (DSA) Plattformen dazu verpflichtet, Meldesysteme für Inhalte einzurichten, Meldungen vertrauenswürdiger Hinweisgeber prioritär zu behandeln, die Rechte der von Meldungen betroffenen Nutzerinnen und Nutzer besser zu wahren und regelmässig Berichte über Risiken und getroffene Massnahmen zu erstellen, ist das in der Schweiz bisher nicht der Fall. Deshalb hat der Bundesrat das Departement für Umwelt, Verkehr, Energie und Kommunikation (UVEK) beauftragt, bis Ende März 2024 eine Vernehmlassungsvorlage auszuarbeiten, mit welcher entsprechende Vorgaben (Meldeverfahren, Begründung für Löschung und Sperrung, Widerspruchsmöglichkeiten, Transparenz und Berichtspflichten) auch in der Schweiz implementiert werden sollen. Aus unserer Sicht sollte eine Plattformregulierung folgende Punkte beinhalten:

- Verpflichtung zur Verbesserung der Kooperation von Plattformbetreibenden mit Strafverfolgungsbehörden,
- Verpflichtung von Plattformbetreibenden zur Einrichtung eines Meldesystems,
- Verpflichtung zur Löschung oder Sperrung von gemeldeten Deepfakes bei begründetem Verdacht auf eine Rechtsverletzung,
- Regelung der Frage, ob Löschungen durch Menschen, automatisiert oder teilautomatisiert unter Hinzuziehung eines Menschen vorgenommen werden sollten (human-in-the-loop-Verfahren),
- Schaffung von mehr Transparenz über Lösch- und Sperrentscheidungen,
- Regelung der Widerspruchsmöglichkeiten für Betroffene,
- Einführung von Berichtspflichten über die Einhaltung der Vorgaben.

### 8.1.2. Unterstützung von Opferberatungsstellen, die auf Cyberdelikte spezialisiert sind

Die prozessuale Durchsetzung von Rechtsansprüchen gegen rechtswidrige Deepfakes ist für Einzelpersonen mit grossem Aufwand verbunden, zudem mangelt es Opfern in der Regel an rechtlichem Fachwissen zur Führung von Rechtsverfahren. Die Unterstützung der Opfer durch spezialisierte Fachstellen könnte insbesondere dort Abhilfe verschaffen, wo es sich nicht um Offizialdelikte handelt (bei denen die Strafverfolgungsbehörden von sich aus tätig werden müssen). Im Bereich der Cybergewalt gegen Frauen ist dies auch völkerrechtlich vorgeschrieben (Art. 20, 21 Istanbul-Konvention). Auf Cyberdelikte spezialisierte Opferberatungsstellen sollten von Bund und Kantonen im Rahmen ihrer Kompetenzen ausreichend personell und finanziell ausgestattet werden.

### 8.1.3. Regelung von digitalen Beweisen im Strafverfahrensrecht (Deepfakes zwecks Visualisierung von Tathergängen oder zur virtuellen Tatortbegehung)

Deepfake-Technologie kann auch zu unschädlichen Zwecken in Rechtsverfahren eingesetzt werden. Genau wie bei anderen digitalen Beweisen (z.B. Virtualreality-Tatortbegehungen) sind aber viele sich damit ergebende Rechtsfragen noch ungeklärt, namentlich die Wahrung der Beschuldigten- und Opferrechte (z.B. Teilnahmerechte, Anbieten von Gegenbeweisen), aber auch die praktische Ablage von digitalen Beweisen in (analogen) Verfahrensdossiers. Diese gilt es zu regeln, um eine Schmälerung der Rechte der Beteiligten zu verhindern und weiterhin eine bestmögliche objektive Sachverhaltsaufklärung zu ermöglichen.

### 8.1.4. Unterstützung vertrauenswürdiger Hinweisgeber (Trusted Flaggers)

Die Schweiz schafft bei Einführung einer Plattformregulierung ähnlich wie in Art. 22 DSA die Möglichkeit von vertrauenswürdigen Hinweisgebern (Trusted Flaggers), die staatlich anerkannt werden und deren Meldungen problematischer Inhalte (nicht nur Deepfakes) von Plattformen prioritär behandelt werden müssen. Gegebenenfalls könnte über eine finanzielle Unterstützung dieser anerkannten vertrauenswürdigen Hinweisgeber nachgedacht werden.

### 8.1.5. Beteiligung an der Schaffung internationaler Normen und Regeln in den Bereichen Desinformation und Cyberkriminalität

Eine nachhaltige Lösung der Probleme in den Bereichen staatlich betriebener, unterstützter oder geduldeter Desinformation und Cyberkriminalität erfordert internationale diplomatische Bemühungen mit dem Ziel, dass diese Staaten die entsprechenden Praktiken reduzieren oder einstellen. Eine Möglichkeit wäre die Erarbeitung internationaler Normen und Regeln zur gegenseitigen medialen Beeinflussung und Einhegung von Cyberkriminalität auf UN-Ebene. Angesichts der zunehmend konfrontativen Weltpolitik können derartige Schritte auch auf Widerstand stossen. Die Alternative wäre jedoch die Inkaufnahme von Desinformationskampagnen und der weiteren Verschlechterung der Cybersicherheit. Dies sollte trotz der wahrscheinlich schwierigen Umsetzbarkeit Anlass genug sein, stärker auf diplomatische Bemühungen zu setzen. Die Schweiz sollte sich, auch in ihrem eigenen Interesse, an solchen Bemühungen beteiligen und sie aktiv unterstützen.

## 8.2. Gesellschaft und Bildungseinrichtungen

Aktuell hat die Schweizer Bevölkerung noch wenig Erfahrungen mit Deepfakes. Nur etwa die Hälfte der Befragten kennt den Begriff Deepfake oder hat schon Deepfakes gesehen. Lediglich eine kleine Minderheit hat schon Deepfakes selbst hergestellt oder weiterverbreitet. Das bedeutet einerseits, dass das Thema Deepfakes durch Unsicherheiten geprägt ist und viele Einschätzungen, Meinungen und Einstellungen zu Deepfakes auf Wahrnehmungen – beispielsweise über Diskurse in Medien – und nicht auf eigenen Erfahrungen beruhen. Andererseits stellt dies auch ein Opportunitätsfenster dar, da mit Bildungs- und Informationsangeboten zur Thematik ein sinnvoller Umgang mit Deepfake-Technologie erlernt werden kann.

### 8.2.1. Sensibilisierungsarbeit in der Ausbildung, Informationskampagnen von staatlichen Akteuren und Engagement journalistischer Medien

Viele Menschen in der Schweiz verbinden Deepfakes überwiegend mit Risiken. Informationsangebote zum Thema Deepfakes an Schulen und über journalistische Medien würden zur Sensibilisierung für Risiken, aber auch Chancen von

Deepfake-Technologien beitragen. Schulen sollten prüfen, ob die Vermittlung von Wissen über Deepfakes unter die Ziele des Lehrplans 21 zur Stärkung der Medienkompetenz fallen, und entsprechend Zeit und Ressourcen im Unterricht einplanen sowie Lehrmittel bereitstellen. Der allgemeinen Medienkompetenz, beispielsweise das Prüfen von Quelle und Absendern oder der Plausibilisierung der Inhalte, kommt nach wie vor eine besondere Bedeutung zu, weil gut gemachte Deepfakes zunehmend schwieriger von realen Inhalten unterschieden werden können. Unsere Studie zeigt, dass dabei auch Kompetenzen im Umgang mit neuen Medien, insbesondere Social Media, zentral sind. Möglich wäre beispielsweise eine Sensibilisierung der Bevölkerung für Chancen und Risiken von Deepfakes über Angebote des öffentlichen Rundfunks, aber auch privaten Medienangeboten für verschiedene Zielgruppen (z.B. auch bildungsfernere Bevölkerungsgruppen und ältere Menschen mit wenig Onlineerfahrung).

### 8.2.2. Selbstverantwortung der Bürgerinnen und Bürger

Es braucht **Selbstverantwortung der Bürgerinnen und Bürger**, um die Bildungsangebote und Aufklärungsarbeiten verschiedener Stellen (bspw. Bildung und Journalismus) wahrzunehmen und zu nutzen. Eine gewisse Selbstverantwortung gilt auch beim Umgang mit Information (Bewertung, Weiterverbreitung und Herstellung von Deepfakes). Das Hochladen von Bildern und Sprachaufnahmen kann die Erstellung von Deepfakes begünstigen (sprich, je mehr audiovisuelles Material von Leuten verfügbar ist, desto eher kann es für Deepfakes missbraucht werden). Deshalb gelten Grundsätze der Medienkompetenz, wie z.B. «das Internet vergisst nicht», und die Reflexion darüber, welche Bilder und Informationen man veröffentlicht, auch in Zusammenhang mit Deepfakes.

## 8.3. Organisationen in allen Branchen

### 8.3.1. Weiterbildungen zu Medien- und Informationskompetenz in sämtlichen Branchen

Bildung bedeutet nicht nur Schulbildung, sondern umfasst auch die **betrieblichen/organisationsinternen Aus- und Weiterbildungen in sämtlichen Branchen**. Die Förderung der erforderlichen Kompetenzen, insb. im Bereich der Medien- und Informationskompetenz zum Umgang mit technischen Innovationen, muss als lebenslanger Prozess verstanden werden, der auch nach der

obligatorischen Schule weiter gefördert werden sollte. Die **Sensibilisierung für (digitale) Technologien und Ethik** sollte insbesondere in der berufsbezogenen Aus- und Weiterbildung und in Studiengängen wie Informatik, wo Personen ausgebildet werden, die Algorithmen und Software entwickeln (die auch zur Erstellung von Deepfakes genutzt werden könnte), stattfinden. Dabei soll ein besonderes Gewicht auf ethische Fragen, Verzerrungen und Menschenrechte im Design von Technologien gelegt werden. Dort, wo Weiterbildung potenzielle Betroffene und Geschädigte von Deepfakes adressiert, müsste über Angriffswege, Erkennungsmöglichkeiten von Deepfakes, Schutzmassnahmen und individuelle Rechte aufgeklärt werden.

### 8.3.2. Förderung von Authentifizierungs- und Kennzeichnungsverfahren

Grundsätzlich werden im Bereich Authentifizierung und Labelling zwei Methoden diskutiert. Zum einen geht es um die Authentifizierung vertrauenswürdiger Inhalte, wie sie etwa seitens der Industrie und grossen Medienorganisationen in Gestalt der Content-Authenticity-Initiative vorangetrieben wird. Zum anderen geht es um die Kennzeichnung von Deepfakes durch Softwarehersteller. Beide Massnahmen sollen eine bessere Unterscheidbarkeit von originalen und synthetischen bzw. manipulierten Inhalten ermöglichen. Beide Methoden werden zugleich hinsichtlich ihrer Wirksamkeit im Kampf gegen irreführende Deepfakes kritisiert.

- Massnahmen der Authentifizierung stehen in der Kritik, weil eine flächendeckende Implementierung trotz grösserer Bemühungen der Industrie infrage steht. Im Missbrauchsfalle könnte ausserdem die Vertrauenswürdigkeit gefälschter Inhalte erhöht werden.
- Gegen Kennzeichnungspflichten spricht erstens, dass Softwarehersteller diese schlicht nicht beachten könnten. Zudem wären solche Hersteller nicht rechtlich fassbar, wenn sie keine Niederlassung in der Schweiz haben. Zweitens würde eine lückenhafte Umsetzung der Labeling-Pflicht darin resultieren, dass sowohl gelabelte als auch ungelabelte Deepfake-Inhalte zirkulieren. So könnte bei ungelabelten Deepfakes in der Bevölkerung der falsche Eindruck entstehen, dass es sich dabei – aufgrund der fehlenden Kennzeichnung – um keine Deepfakes handelt. Ohne es zu wollen, könnte die Kennzeichnungspflicht somit die Glaubwürdigkeit von Deepfakes steigern.

Eine umfassende Lösung ist insofern weder von Authentifizierungs- noch von Kennzeichnungsverfahren zu erwarten. Dennoch können beide Methoden ein hilfreiches Element bei der Erhaltung des Vertrauens in Medieninhalte darstellen, indem bei Umsetzung der Massnahmen ein Grossteil originaler Inhalte bzw. Deepfakes als solche gekennzeichnet wären. Nicht gekennzeichnete bzw. authentifizierte Deepfakes müssten dann mittels über normale Vertriebswege nicht erhältlicher Software oder fachkundiger Programmierung erstellt werden. Das könnte den Kreis der Personen, die eine missbräuchliche Nutzung durchführen können, noch stärker auf technisch versierte bzw. ressourcenstarke Angreifer und damit auch die absolute Zahl nicht gekennzeichneter Deepfakes deutlich begrenzen.

Daher können wir zwar keine uneingeschränkte Empfehlung zur Nutzung oder verpflichtenden Einführung der Methoden abgeben, wir halten es jedoch für sinnvoll, dass sowohl die Politik als auch etwaige Organisationen, die ein Interesse an der Authentifizierung ihrer Inhalte haben, beide Diskussionsstränge weiterverfolgen und ggf. mit Fördermassnahmen eine Umsetzung ausprobieren und unterstützen (vgl. auch Kapitel 2.6.1.1).

### 8.3.3. Nutzung fortschrittlicher Authentifizierungsverfahren und von Zwei-Faktor-Authentifizierung

Schon heute existieren Möglichkeiten der computerbasierten originalgetreuen Imitation einer Stimme oder eines Gesichts zur Überwindung biometrischer Authentifizierungsmassnahmen. Wenn z.B. auf die Stimme als alleinigen Authentifizierungsfaktor gesetzt wird, könnten Angreifer Zugang zu Bankkonten, sensiblen Daten usw. erhalten. Dies kann sowohl zu Identitätsdiebstahl bei Privatpersonen führen als auch zu Identitätsbetrug gegenüber Unternehmen. Entwickler von Authentifizierungstechnologien sind sich über diese Schwächen im Klaren und forschen an Möglichkeiten der Authentifizierung unter Hinzuziehung weiterer Merkmale. Dadurch sollen Täuschungsversuche mittels Stimm- oder Gesichts-Deepfakes vereitelt werden. Die Entwickler von Deepfake-Technologien arbeiten derweil an der Überlistung dieser modifizierten Systeme, etwa indem die Lebenderkennung bei gesichtsbiometrischen Systemen überlistet wird. Im Ergebnis findet auch in diesem Bereich ein Wettrennen zwischen Angreifern und Verteidigern statt. Wir begrüssen die Bemühungen der Entwickler zum Schutz vor Deepfakes, warnen zugleich jedoch vor der Gefahr und empfehlen, bei derartigen Systemen auf fortschrittliche Authentifizierungsverfahren und Zwei-Faktor-Authentifizierung zu setzen.

### 8.3.4. Freiwillige Meldung von Deepfake-Vorfällen durch betroffene Organisationen

Nach Bekanntwerden des ersten Deepfake-Voice-Phishing-Falls wurde vielfach über dieses Thema berichtet und das Treffen von geeigneten Schutzmassnahmen zur Verhinderung vergleichbarer Angriffe bei anderen Unternehmen war eins der prioritär behandelten Aspekte. Das Melden eines Angriffs kann zunächst dem betroffenen Unternehmen bei der Schadensbegrenzung helfen, etwa indem die zuständigen Behörden Hilfestellungen geben. Das Bekanntwerden von Angriffsvektoren kann ausserdem bei anderen Organisationen zu erhöhter Aufmerksamkeit, der Verbesserung von Schutzmassnahmen und damit einer verringerten Chance des Erfolgs eines Angriffs führen. Als Meldestelle kommen neben den zuständigen Polizeistellen auch der BACS oder private Beratungsagenturen infrage.

### 8.3.5. Einrichtung von spezialisierten Teams, die im Falle eines Deepfake-Einsatzes darauf vorbereitet sind, Massnahmen zur Schadensbegrenzung zu ergreifen

Deepfakes, die gefälschte Aussagen oder Handlungen der abgebildeten Personen (z.B. Politikerinnen und Politiker oder Führungspersonen eines Unternehmens) zum Gegenstand und z.B. Rufschädigung von einer Person oder Organisation zum Ziel haben, können wirkungsvoller sein, wenn die Zielperson oder Organisationen träge in ihrer Reaktion sind. Daher kann es ratsam sein, dass Organisationen spezialisierte Stellen einrichten bzw. bestehendes Personal (z.B. Presse- und Kommunikationsverantwortliche) entsprechend schulen, die im Falle eines Deepfake-Einsatzes in der Lage sind, Massnahmen zur Schadensbegrenzung zu ergreifen. Hierzu zählt z.B. die schnelle Aufnahme und Verbreitung von audiovisuellen Klarstellungen (für weitere Details dieser Empfehlung vgl. Unterkapitel 7.4.2.1).

## 8.4. Kommunikationsbranche

### 8.4.1. Selbstregulierung der PR- und Werbebranche

Für in der Kommunikationsbranche tätige Akteure wie Werbe-, Kommunikations- und PR-Agenturen bieten Deepfakes neue Möglichkeiten für die (visuelle) Kommunikation. Das gilt auch für Influencerinnen und Influencer auf sozialen Netzwerken und Video-Sharing-Plattformen und deren Marketingdienstleistun-

gen. Deepfakes können durchaus kreativ und gefahrlos eingesetzt werden. Allerdings besteht auch die Gefahr, dass Auftraggeber aus Wirtschaft, Politik und Verwaltung sowie dem dritten Sektor Deepfakes auf intransparente Weise oder mit Täuschungsabsicht einsetzen möchten. Insbesondere mit Blick auf politische Kommunikation und PR (bspw. Wahl- und Abstimmungskampagnen) erscheint dies potenziell problematisch. Im eidgenössischen Wahlkampf 2023 kamen bereits vereinzelt KI-generierte Bilder und Videos zum Einsatz und die Parteien präsentieren sich uneins über den Umgang mit Deepfakes. Entsprechend besteht Bedarf nach einer (aktualisierten) Selbstregulierung der Kommunikationsbranche zum Umgang mit Deepfakes. Gefordert sind hierbei bestehende Akteure wie die Schweizerische Lauterkeitskommission (SLK), die Schweizerische Public Affairs Gesellschaft (SPAG) und der Schweizerische Public Relations Verband (SPRV), aber auch neue Initiativen wie der Code of Conduct des «Conscious Influence Hub» (https://www.consciousinfluencehub.org/our-mission). Neben Akteuren in der Kommunikationsbranche sind aber auch deren Auftraggeber gefordert. Mehrere Parteien haben sich zwischenzeitlich auch auf gemeinsame Regeln für den Einsatz von KI im Wahlkampf geeinigt.

## 8.5. Plattformbetreiber

### 8.5.1. Selbstregulierungsmassnahmen gegen irreführende und illegale Inhalte

Der Gesetzgeber sollte Massnahmen der Selbstregulierung der Plattformen einfordern und fördern, ähnlich der Vorgabe im Digital Services Act. Plattformbetreiber sollten deshalb auch weiterhin selbstständig Massnahmen treffen, um die schädigende Wirkung von potenziell irreführenden und illegalen Inhalten zu reduzieren. Und auch dann, wenn der gegenwärtige Vorstoss zur Plattformregulierung erfolgreich sein sollte, könnten Plattformbetreiber freiwillig über die staatlichen Vorgaben hinausgehen. Hierzu zählt zumindest die Zusammenarbeit mit Faktencheckern bzw. Trusted Flaggers, aber auch die Einrichtung eines Meldesystems, die Herstellung von Transparenz über Lösch- und Sperrentscheidungen und von Widerspruchsmöglichkeiten für Betroffene sowie die regelmässige Berichterstattung über die Umsetzung derartiger Massnahmen. Dazu zählen auch Massnahmen, die nicht nur die Löschung von gemeldeten Deepfakes bei begründetem Verdacht auf eine Rechtsverletzung beinhalten, sondern auch beim Versuch des erneuten Hochladens eines solchen Inhalts greifen.

## 8.6. Medienorganisationen, Medienausbildung, Nachrichtenagentur

### 8.6.1. Hochhaltung journalistischer Standards bei der Erkennung von Deepfakes und Förderung der Medienethik

Durch den technologischen Fortschritt wird Desinformation noch raffinierter und erscheint authentischer, was deren Erkennung erschwert. Die traditionellen journalistischen Standards spielen hierbei eine zentrale Rolle. Die normorientierte journalistische Arbeit wird zwar in der Ausbildung vermittelt, allerdings beeinflussen viele Faktoren den Arbeitsalltag von Journalistinnen und Journalisten, etwa ökonomische Ziele oder Publikationsdruck.

Die Einhaltung dieser Standards ist auch im Umgang mit Deepfakes zentral. Nur so kann der Journalismus seine gesellschaftliche Funktion – die Bereitstellung von nach professionellen Standards produzierten Inhalten auf Basis geprüfter Informationen – wahrnehmen und sich als vertrauenswürdige Informationsquelle behaupten. Allerdings haben Ressourcenmangel und Zeitdruck durch die anhaltende Finanzierungskrise des Journalismus zugenommen, was sich auch auf die Möglichkeiten von Redaktionen zur Hochhaltung der Standards auswirkt. Entsprechend sollte eine Medienförderung in Betracht gezogen werden.

Die Berücksichtigung journalistischer Normen ist für die Realisierung der Chancen von Deepfakes wichtig. Das Ausprobieren neuer technischer Lösungen ist integraler Bestandteil des Innovationsprozesses in der Medienproduktion. Allerdings sind ethische Überlegungen hinsichtlich der Ziele, des Nutzens und potenzieller Gefahren von Deepfakes entscheidend, um Risiken zu reduzieren. Darauf sollten sich Medien- wie Ausbildungsorganisationen vorbereiten.

### 8.6.2. Förderung forensischer Verifikationsmethoden in den Redaktionen und Herstellung von Transparenz über eigene Bemühungen

Bei der Überprüfung komplexer Informationen – wie etwa der Echtheit von Bild- und Videomaterial – gewinnen die Stärkung von Basiswissen der Medienschaffenden über Open-Source Intelligence (OSINT) und technische Hilfsmittel an Bedeutung. Die Förderung solchen Wissens in der Aus- und Weiterbildung ist deshalb empfehlenswert. Um insbesondere kleinere Medienanbieter finanziell zu entlasten, sind Kooperationen zwischen Medienhäusern empfehlenswert.

Solche sind etwa in der OSINT-Reporter-Community bereits erkennbar, wo sich Mitglieder gegenseitig bei der Informationsüberprüfung unterstützen. Zur Realisierung von Synergien wäre auch die Stärkung eines entsprechenden Kompetenzteams bei der Nachrichtenagentur Keystone-SDA sinnvoll, wovon die ganze Medienbranche profitieren würde. Insbesondere kleineren Medienanbietern würde dies bessere Möglichkeiten für die Erkennung von Deepfakes eröffnen. Angesichts von Deepfakes und anderen Formen der Desinformation wäre auch eine Berichterstattung über eigene Verifikationsprozesse sinnvoll, um journalistische Arbeitsprozesse verständlich zu machen, die eigenen Verifikationsanstrengungen aufzuzeigen und das Vertrauen in den Journalismus zu fördern. Bessere Einblicke in die Medienproduktion könnten zudem einen Beitrag zur Stärkung der Medien- und Digitalkompetenz des Publikums und damit zum kritischen Umgang mit Informationen auf Onlineplattformen leisten.

### 8.6.3. Stärkung des Presserats als von der Branche eingesetzte Selbstregulierungsorganisation zur Einhaltung ethischer Standards im Journalismus

Die Digitalisierung bringt für die journalistische Arbeit neue Herausforderungen mit sich, auf die der Presserat reagieren muss, beispielsweise durch eine Anpassung seiner Richtlinien. Dies ist auch denkbar mit Blick auf die Manipulation von Bildern, Ton und Videos. Der Presserat könnte so wichtige Impulse für den Umgang mit synthetischen Inhalten und Deepfakes im Schweizer Journalismus setzen. Um seine wichtige Aufgabe weiterhin wahrnehmen zu können und darüber hinaus auf die digitale Entwicklung reagieren zu können, scheint eine grössere finanzielle Unterstützung unverzichtbar zu sein.

# 9. Schlussfolgerungen

*Murat Karaboga*

Neue Technologien zur Synthetisierung und Manipulation von Medien werden künftig einen festen Platz in der Alltagskultur einnehmen. Insofern wird in der vorliegenden Studie der Umgang mit einer veränderten Realität diskutiert, in der täuschend echte Fälschungen mit originalen Medieninhalten um die Gunst der Rezipienten buhlen.

Dabei wird zunächst deutlich, dass das ganze Feld durch eine hohe Entwicklungsdynamik geprägt ist. Allein während der etwa zweijährigen Entstehungszeit dieser Studie hat sich der Diskurs über Deepfakes erheblich gewandelt. Durch die Veröffentlichung von KI-basierten Bild-, Audio- und Textgeneratoren hat sich die Debatte vom Fokus auf die Manipulation des Aussehens und Ausdrucks von Personen gelöst. Stattdessen treten neue Anwendungsmöglichkeiten in den Vordergrund, die von der Erschaffung neuer Welten mittels Bild- und Videogeneratoren bis hin zur Unterstützung mittels LLM-basierter Chatbots in Beruf und Alltag reichen. Allerdings muss festgehalten werden, dass ein grosser Teil der Menschen in der Schweiz noch keine oder nur wenig Erfahrung mit Deepfakes haben. Viele Einschätzungen, Meinungen und Einstellungen zu Deepfakes basieren auf Wahrnehmungen – bspw. über Diskurse in Medien – und nicht auf eigenen Erfahrungen.

Dabei zeigt sich zugleich, dass sich bisherige Schreckensvisionen von einer «Überflutung» der Medienlandschaft mit gefälschten Inhalten und der Erosion des Vertrauens in Nachrichten, Politik usw. – bislang – nicht bewahrheitet haben. Einen Anteil daran hat – neben anhaltenden medienpädagogischen Aufklärungsbemühungen – sicherlich, dass fortschrittliche Deepfake-Software noch immer nicht von Laien zur Produktion hochwertiger Fälschungen genutzt werden kann. Wir gehen sogar davon aus, dass kurzfristig technische Laien und damit die grosse Mehrheit der Bevölkerung nicht in der Lage sein werden, täuschend echte Deepfake-Videos zu erstellen. Über einfache Wege (z.B. App Stores) erwerbbare Anwendungen bieten weiterhin lediglich die Erstellung von eher durchschnittlichen Deepfake-Videos an, die oft mit Wasserzeichen versehen sind oder bereits aufgrund der geringen Qualität und aus dem Kontext heraus als Deepfake erkannt werden, etwa wenn die Gesichter von Schauspielerinnen und Schauspielern aus bekannten Filmen und Serien mit dem eigenen Gesicht ersetzt werden. Akteure, die über das nötige Know-how und Ressourcen verfügen, können allerdings schon heute hochwertige Deepfakes erstellen.

Deepfakes und synthetische Medieninhalte werden sich voraussichtlich weiterhin in die kulturellen Gewohnheiten der Gesellschaft einfügen – im Guten wie im Schlechten. Im Schlechten etwa, indem sich Deepfakes und synthetische Medien voraussichtlich weiter in das Angriffsrepertoire in den Bereichen Desinformation und Cyberkriminalität einfügen werden. Im Guten hingegen werden Deepfakes und synthetische Medien kulturelle und wirtschaftliche Innovationen ermöglichen.

Neben der Bestandsaufnahme des gegenwärtigen Stands von Deepfake-Technologien bzw. von KI-basierten Audio-, Text- und Bildgeneratoren im Rahmen einer Ist- und Trendanalyse beinhaltet die Studie die vertiefte Untersuchung von fünf Problemkomplexen im Bereich Deepfakes und manipulierter Medien. Der Fokus liegt dabei auf bildbasierten Deepfakes. Wo es sich anbietet, werden auch Deepfake-Audios und KI-generierte Texte mitbetrachtet.

In der ersten inhaltlichen Vertiefung wurde im Rahmen einer Bevölkerungsumfrage untersucht, wie die Wahrnehmung von Deepfakes in der Schweizer Bevölkerung ist. Es zeigte sich, dass fast die Hälfte der Befragten den Begriff Deepfake nicht kennt. Es besteht folglich noch wenig Erfahrung im Umgang mit Deepfakes in der Bevölkerung. Die Untersuchung zeigt weiter, dass Menschen kaum in der Lage sind, gut gemachte Deepfakes als solche zu erkennen. Positiv auf die Erkennungskompetenz wirkt sich die Kompetenz im Umgang mit Social Media aus. Deepfake-Technologie wird zudem viel stärker mit Risiken als mit Chancen assoziiert, insbesondere für Politik und Medien, während für die Wirtschaft die Risiken etwas geringer eingeschätzt werden. Verwendet man anstelle des Begriffs Deepfake die Bezeichnung «synthetische Medien», werden eher Chancen assoziiert. Das Framing und die Begrifflichkeiten in der Diskussion zu Deepfake-Technologien sind somit im Hinblick auf gesellschaftliche Akzeptanz und Ablehnung von Bedeutung.

Die Untersuchung zu den Herausforderungen von Deepfakes für den Schweizer Journalismus zeigt, dass dort Deepfakes als technischer Sonderfall von Desinformation behandelt werden und als solcher Eingang in journalistische Verifikationsprozesse sowie die Schulung von Journalistinnen und Journalisten findet. Im Umgang mit Deepfakes werden keine auf diese zugeschnittenen Massnahmen ergriffen, sondern bestehende journalistische Normen und Standards hochgehalten. Es zeigten sich einige Bestrebungen von Redaktionen, Verifikationsteams auf- und auszubauen, um den heutigen und zukünftigen Herausforderungen bei der Verifikation von komplexen Informationen entgegenzutreten.

Die Analyse von Deepfakes im Recht verdeutlicht, dass Deepfakes in der Schweiz derzeit keiner spezifischen Regulierung, sondern – je nach Anwendung – unterschiedlichen allgemeinen Rechtsvorschriften unterliegen. Neben zivil- und strafrechtlichen Bestimmungen sind das öffentlich-rechtliche Vorgaben im Rundfunkrecht und im Bereich politischer Rechte, die zur Prävention missbräuchlicher Einsätze von Deepfakes dienen können.

Die Untersuchung von Deepfakes in der Wirtschaft und in der Politik zeigt insbesondere, wie Deepfakes zu Zwecken der Desinformation und Cyberkriminalität eingesetzt werden könnten und welche Handlungsoptionen sich zur Abwehr und Verhinderung derartiger Angriffe eignen. Deutlich wurde, dass Akteure aus Politik, Wirtschaft und Medien schon heute institutionelle und persönliche Massnahmen zur Prävention und Abwehr von schädlichen Deepfakes treffen können. Hierzu zählt die Schaffung neuer oder Schulung bestehender Reaktionsteams, die schnelle Verbreitung von Richtigstellungen durch Medienorganisationen, die Sensibilisierung von Mitarbeitenden für verdächtige Situationen sowie das Treffen von grundlegenden IT-Sicherheitsvorkehrungen.

Eine im Rahmen der Studie durchgeführte Umfrage unter Schweizer Parlamentarierinnen und Parlamentariern zum Thema Deepfakes zeigte, dass die Befragten fast ausschliesslich Risiken durch Deepfakes sehen, insb. Gefahren für die Schweizer Demokratie und politischen Institutionen. Zugleich wurde die Gefahr, selbst Opfer eines Deepfakes zu werden, im Vergleich zu den Gefahren für das politische System als gering eingeschätzt. Schliesslich zeigte sich bei den Befragten eine dahin gehende Tendenz, dass gegenwärtig keine konkreten Schutzmassnahmen gegen Deepfakes ergriffen werden.

Wir zeigen schliesslich Handlungsmöglichkeiten für die Politik, Medienorganisationen, Bildungsinstitutionen, Plattformbetreiber und weitere Akteure auf. Insbesondere betreffen die Empfehlungen die Bereiche Plattformregulierung, Bildung und Selbstverantwortung der Bürgerinnen und Bürger, Vorbereitung auf Deepfakes in Organisationen sowie das Hochhalten journalistischer Standards.

Einen wichtigen kommenden Entwicklungsschritt bei Deepfake-Technologien stellt die derzeit noch laufende Entwicklung von Text-to-Video-Technologien dar. Heutige Deepfake-Software zur Erschaffung hochwertiger Deepfakes erfordert technisches Know-how, vergleichsweise grosse Rechenkapazitäten und damit Geld, Zeit und Vorbereitung im Vorfeld, etwa die Suche nach geeigneten Bild- und Tonaufnahmen, die für das Training der Deepfake-Software erforderlich sind. Mit Text-to-Video-Technologien wird sich dieser Erstellungsprozess voraussichtlich enorm vereinfachen. Ähnlich wie bei den bereits verfügbaren Bild-,

Audio- und Textgeneratoren werden Nutzende Deepfake-Videos dann mit einfachen Texteingaben erstellen können. Diese Entwicklung einfach erstellbarer Deepfakes sollte kritisch beobachtet werden. Handlungsbedarf ergibt sich aber bereits schon heute, sodass schädlichen Nutzungen von Deepfakes besser vorgebeugt und etwaige unerwünschte Auswirkungen reduziert werden. Denn Anwendungen von Deepfakes können auch dann schädliche Effekte nach sich ziehen, wenn die Erschaffer von Inhalten und Technologien gar keine Schädigungsabsicht verfolgen. Besonders wird dies an den Debatten über die Auswirkungen der Nutzung von KI-Generatoren deutlich.

# Autorinnen und Autoren

**Murat Karaboga**, Dr. phil. (Studienleiter): Politikwissenschaftler und Projektleiter im Geschäftsfeld Informations- und Kommunikationstechnologien (IKT) am Competence Center (CC) Neue Technologien des Fraunhofer ISI. Lehrbeauftragter am Institut für Politikwissenschaft der Johann Wolfgang-Goethe-Universität Frankfurt am Main.

**Frank Ebbers**, Dr. rer. pol.: Wirtschaftsinformatiker und Post-Doc im Geschäftsfeld IKT am CC Neue Technologien am Fraunhofer ISI.

**Michael Friedewald**, Dr. Ing. (Co-Studienleiter): Ingenieur der Elektrotechnik und Informationstechnik sowie Wirtschaftswissenschaftler. Koordinator des Geschäftsfelds IKT und Senior-Researcher im CC Neue Technologie am Fraunhofer ISI.

**Greta Runge**, M. A.: Politikwissenschaftlerin und Doktorandin im Geschäftsfeld IKT am CC Neue Technologien am Fraunhofer ISI.

**Nula Frei**, Ass.-Prof. Dr. iur.: Assistenzprofessorin an der FernUni Schweiz; ehemalige Lehr- und Forschungsrätin am Institut für Europarecht der Universität Freiburg i. Ue.

An einer früheren Fassung dieser Studie hat auch Prof. Dr. habil. **Benedikt Pirker**, Titularprofessor am Institut für Europarecht der Universität Freiburg i. Ue., mitgearbeitet.

**Sophia Rovelli**, MLaw, Doktorandin; ehemalige wissenschaftliche Mitarbeiterin am Institut für Europarecht der Universität Freiburg i. Ue.

**Manuel Puppis**, Prof. Dr.: Ordentlicher Professor am Departement für Kommunikationswissenschaft und Medienforschung (DCM) der Universität Freiburg i. Ue.

**Patric Raemy**, Dr. rer. soc.: Oberassistent am Departement für Kommunikationswissenschaft und Medienforschung (DCM) der Universität Freiburg i. Ue.

**Gwendolyn Gurr**, Dr. rer. soc.: ehemalige Oberassistentin am Departement für Kommunikationswissenschaft und Medienforschung (DCM) der Universität Freiburg i. Ue.

**Daniel Vogler**, Dr.: stv. Direktor Forschungszentrum für Öffentlichkeit und Gesellschaft (fög), Universität Zürich und Oberassistent am Institut für Kommunikations- und Medienforschung, Universität Zürich (IKMZ).

**Adrian Rauchfleisch**, Prof. Dr.: Associate Professor am Graduate Institute of Journalismus an der National Taiwan University.

**Gabriele de Seta**, (PhD): Postdoktorand am Department für Linguistik, Literatur and ästhetische Studien an der Universität Bergen (NO).

# Literatur


AAP (2020), WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): WIPO/IP/AI/2/GE/20/1.

Adobe (2022), Content Authenticity Initiative, 13.05.2022, https://contentauthenticity.org/.

Aebi-Müller, Regina E. (2005), *Personenbezogene Informationen im System des zivilrechtlichen Persönlichkeitsschutzes*. *Unter besonderer Berücksichtigung der Rechtslage in der Schweiz und in Deutschland*, Bern: Stämpfli.

Agarwal, Shruti/Hany Farid/Tarek El-Gaaly/Ser-Nam Lim (2020), Detecting Deep-Fake Videos from Appearance and Behavior, in: *IEEE WIFS*: IEEE, S. 1–6.

Agarwal, Shruti/Hany Farid/Yuming Gu/Mingming He/Koki Nagano/Hao Li (2019), Protecting World Leaders Against Deep Fakes, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Ahmed, Saifuddin (2021a), Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism, *New Media & Society*, 1–22.

Ahmed, Saifuddin (2021b), Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size, *Telematics and Informatics*, Jg. 57, S. 101508.

Ajder, Henry/Giorgio Patrini/Francesco Cavalli/Laurence Cullen (2019), The state of deepfakes: Landscape, threats, and impact, 16.11.2023, https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

Alba, Davey (06.01.2020), George Floyd's 'Fake' Death and Other Misinformation Spread Online, *The New York Times*, https://www.nytimes.com/2020/06/01/technology/george-floyd-misinformation-online.html, 16.11.2023.

Ali, Safinah/Daniella DiPaola/Irene Lee/Victor Sindato/Grace Kim/Ryan Blumofe/Cynthia Breazeal (2021), Children as creators, thinkers and citizens in an AI-driven future, *Computers and Education: Artificial Intelligence*, Jg. 2, S. 100040.

Allen, Jennifer/Baird Howland/Markus Mobius/David Rothschild/Duncan J. Watts (2020), Evaluating the fake news problem at the scale of the information ecosystem, *Science advances*, Jg. 6, H. 14, eaay3539.

Almutairi, Zaynab/Hebah Elgibreen (2022), A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions, *Algorithms*, Jg. 15, H. 5.

Altmeppen, Klaus-Dieter (2006), *Journalismus und Medien als Organisationen. Leistungen, Strukturen und Management,* 1. Aufl., Wiesbaden: Verlag für Sozialwissenschaften.

Amezaga, Naroa/Jeremy Hajek (2022), Availability of Voice Deepfake Technology and its Impact for Good and Evil, in: Ray Trygstad/Yong Zheng (Hg.), *Proceedings of the SIGITE '22*, NY, USA: ACM, S. 23–28.

AP (2008), A New Model for News. Studying the Deep Structure of Young-Adult News Consumption.

Appel, Markus/Fabian Prietzel (2022), The detection of political deepfakes, *Journal of Computer-Mediated Communication*, Jg. 27, H. 4.

Arjovsky, Martin/Léon Bottou (2017), Towards Principled Methods for Training Generative Adversarial Networks, http://arxiv.org/pdf/1701.04862v1.

Aschenbrenner, Vali (2020), The Mandalorian: Deepfake-KI schlägt Disney-CGI um Längen, *GameStar*.

Ashley, Seth/Adam Maksl/Stephanie Craft (2013), Developing a News Media Literacy Scale, *Journalism & Mass Communication Educator*, Jg. 68, H. 1, S. 7–21.

Ayyub, Rana (November 21, 2018), I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me, *Huffington Post*, https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

Aznar, Eduardo (2021), *Who Exactly Is Behind Ad Fraud?*, *blog.opticksecurity.com*, https://blog.opticksecurity.com/who-is-behind-ad-fraud, 15.05.2023.

BAKOM (2011), Rechtliche Basis für Social Media. Bericht des Bundesrates in Erfüllung des Postulats Amherd 11.3912 vom 29.September 2011.

BAKOM (2018), Strategie «Digitale Schweiz», https://www.bakom.admin.ch/dam/bakom/de/dokumente/informationsgesellschaft/strategie2018/strategie%20digitale%20schweiz.pdf.download.pdf/strategie%20digitale%20schweiz%20DE.pdf.

Balla, Steven/Reeve Bull/Bridget Dooling/Emily Hammond/Michael Livermore/Michael Herz/Beth Noveck (2022), Responding to Mass, Computer-Generated, and Malattributed Comments, *Articles*, Jg. 74, S. 95.

Bao, Luye/Nicole M. Krause/Mikhaila N. Calice/Dietram A. Scheufele/Christopher D. Wirz/Dominique Brossard/Todd P. Newman/Michael A. Xenos (2022), Whose AI? How different publics think about AI and its social impacts, *Computers in Human Behavior*, Jg. 130, S. 107182.

Bareis, Jascha/Christian Katzenbach (2022), Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics, *Science, Technology, & Human Values*, Jg. 47, H. 5, S. 855–881.

Barrelet, Denis/Willi Egloff/Michel Heinzmann/Sandra Künzi/Dieter Meier/Christof Riedo (2020), *Das neue Urheberrecht*. *Kommentar zum Bundesgesetz über das Urheberrecht und verwandte Schutzrechte,* 4. Auflage, vollständig überarbeitet und ergänzt, Bern: Stämpfli Verlag.

Bartz, Tim (10.05.2022), Drama um die Credit Suisse: Droht in der Schweiz ein neues Lehman Brothers?, *DER SPIEGEL*, https://www.spiegel.de/wirtschaft/unternehmen/drama-um-die-credit-suisse-droht-in-der-schweiz-ein-neues-lehman-brothers-a-16da659b-eea8-4be9-b06f-2061150593e0, 17.05.2023.

Barua, Sukarna/Xingjun Ma/Sarah Monazam Erfani/Michael E. Houle/James Bailey (2019), Quality Evaluation of GANs Using Cross Local Intrinsic Dimensionality.

Bastian, Matthias (2021), *Mehr Retro Star Wars? Lucasfilm schnappt sich Deepfaker*, *MIXED*, https://mixed.de/mehr-retro-star-wars-lucasfilm-schnappt-sich-youtube-deepfaker/, 02.06.2022.

Bateman, Jon (2020), Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios, *Cyber Policy Initiative Working Paper Series*.

Baumann, Sophia/Christo Buschek/Maria Christoph/Max Hoppenstedt/Carina Huppertz/Dajana Kollig/Hannes Munzinger/Frederik Obermaier/Bastian Obermayer/Marcel Rosenbach/Hakan Tanriverdi (3/30/2023), Massiver Leak legt erstmals Russlands Krieg im Netz offen, *DER STANDARD*, https://www.derstandard.at/story/2000145052847/massiver-leak-legt-erstmals-russlands-krieg-im-netz-offen, 16.05.2023.

Baur, Nina/Jörg Blasius (2019), *Handbuch Methoden der empirischen Sozialforschung*, Wiesbaden: Springer Fachmedien Wiesbaden.

Baxevanakis, Spyridon/Giorgos Kordopatis-Zilos/Panagiotis Galopoulos/Lazaros Apostolidis/Killian Levacher/Ipek B. Schlicht/Denis Teyssou/Ioannis Kompatsiaris/Symeon Papadopoulos (2022), The MeVer DeepFake Detection Service: Lessons Learnt from Developing and Deploying in the Wild.

Beranek Zanon, Nicole (2022), Deepfakes – Neue rechtliche Herausforderungen aufgrund technologischen Fortschritts, 04.11.2023, https://haerting.ch/wissen/deepfakes/.

Beuth, Patrick (2019), Deepfake-Zuckerberg erklärt sich zum Weltherrscher, *Spiegel Online*.

Bezmalinovic, Tomislav (11/17/2020), «Deepfakes» – Wenn Merkel plötzlich Trumps Gesicht trägt: die gefährliche Manipulation von Bildern und Videos, *Aargauer Zeitung*, https://www.aargauerzeitung.ch/leben/digital/wenn-merkel-plotzlich-trumps-gesicht-tragt-die-gefahrliche-manipulation-von-bildern-und-videos-ld.1481742, 16.11.2023.

Binder, Andrew R./Michael A. Cacciatore/Dietram A. Scheufele/Bret R. Shaw/Elizabeth A. Corley (2012), Measuring risk/benefit perceptions of emerging technologies and their potential impact on communication of public opinion toward science, *Public Understanding of Science*, Jg. 21, H. 7, S. 830–847.

BJ (2015), Die zivilrechtliche Verantwortlichkeit von Providern. Bericht des Bundesrates vom 11. Dezember 2015, 06.11.2023, https://www.bj.admin.ch/bj/de/home/publiservice/publikationen/berichte-gutachten/2015-12-11.html.

Blattmann, Andreas/Robin Rombach/Huan Ling/Tim Dockhorn/Seung Wook Kim/Sanja Fidler/Karsten Kreis (2023), Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models, https://arxiv.org/abs/2304.08818.

Block, Katharina/Sascha Dickel (2020), Jenseits der Autonomie die De/Problematisierung des Subjekts in Zeiten der Digitalisierung, *BEHEMOTH A Journal on Civilisation*, Volume 13 Issue No. 1.

Bogner, Alexander/Michael Decker/Michael Nentwich/Constanze Scherz (2022), 24, in: Valerie Martínez (Hg.), *Each and Her*: University of Arizona Press, S. 24.

Bogner, Alexander/Wolfgang Menz (2002), Das theoriegenerierende Experteninterview. Erkenntnisinteresse, Wissensformen, Interaktion, in: Alexander Bogner/Beate Littig/Wolfgang Menz (Hg.), *Das Experteninterview*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 33–70.

Bolzern, Tobias/Jonas Bucher (3/17/2022), «Legt Waffen nieder» – Falscher Selenski sorgt mit Deepfake für Chaos, *20 Minuten*, https://www.20min.ch/story/legt-waffen-nieder-falscher-selenski-sorgt-fuer-chaos-im-netz-256597207602, 16.11.2023.

Borchardt, Alexandra (2022), Go, Robots, Go! the Value and Challenges of Artificial Intelligence for Local Journalism, *Digital Journalism*, Jg. 10, H. 10, S. 1919–1924.

Borucki, Isabelle/Dennis Michels/Stefan Marschall (2020), Die Zukunft digitalisierter Demokratie – Perspektiven für die Forschung, *Zeitschrift für Politikwissenschaft*, Jg. 30, H. 2, S. 359–378.

Bradshaw, S./P. Howard (2017), Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation, *Computational Propaganda Research Project*.

Brandtzaeg, Petter Bae/Marika Lüders/Jochen Spangenberg/Linda Rath-Wiggins/Asbjørn Følstad (2016), Emerging Journalistic Verification Practices Concerning Social Media, *Journalism Practice*, Jg. 10, H. 3, S. 323–342.

Breithut, Jörg (2015), App Face Swap Live: Gesicht tauschen mit Donald Trump, 01.08.2022, https://www.spiegel.de/netzwelt/apps/app-face-swap-live-gesicht-tauschen-mit-donald-trump-a-1068517.html.

Breitschmid, Peter/Annasofia Kamp, Persönlichkeitsschutz Verstorbener – Urheberpersönlichkeitsschutz im Besonderen, *successio*, Jg. 2011, H. 1, S. 19–29.

Breyer, Patrick (2023), Digitale-Dienste-Gesetz (Archiv), https://www.patrick-breyer.de/beitraege/dsa/.

Brode, Bernard (10.12.2020), Deepfake Voice Technology Iterates on Old Phishing Strategies, *Tripwire*, https://www.tripwire.com/state-of-security/featured/deepfake-voice-technology-phishing-strategies/, 09.10.2022.

Brown, Tom B./Benjamin Mann/Nick Ryder/Melanie Subbiah/Jared Kaplan/Prafulla Dhariwal/Arvind Neelakantan/Pranav Shyam/Girish Sastry/Amanda Askell/Sandhini Agarwal/Ariel Herbert-Voss/Gretchen Krueger/Tom Henighan/Rewon Child/Aditya Ramesh/Daniel M. Ziegler/Jeffrey Wu/Clemens Winter/Christopher Hesse/Mark Chen/Eric Sigler/Mateusz Litwin/Scott Gray/Benjamin Chess/Jack Clark/Christopher Berner/Sam McCandlish/Alec Radford/Ilya Sutskever/Dario Amodei (2020), Language Models are Few-Shot Learners, http://arxiv.org/pdf/2005.14165v4.

BSI (2021b), IT-Sicherheitsleitfaden für Kandidierende bei Bundes- und Landeswahlen, Berlin, 24.04.2023, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Broschueren/Leitfaden-Kandidierende.pdf?__blob=publicationFile&v=7.

BSI (2022a), Deepfakes - Gefahren und Gegenmaßnahmen, 17.11.2022, https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html#doc1009562bodyText6.

BSI (2022b), Ransomware Bedrohungslage 2022, Bonn.

Buchstein, Hubertus (2012), Divergierende Konzepte Politischen Handelns in der Politikwissenschaft, in: Georg Weisseno/Hubertus Buchstein (Hg.), *Politisch Handeln: Modelle, Möglichkeiten, Kompetenzen*: Bundeszentrale für Politische Bildung, S. 18–38.

Budhkar, Akshay/Krishnapriya Vishnubhotla/Safwan Hossain/Frank Rudzicz, Generative Adversarial Networks for Text Using Word2vec Intermediaries, in: Isabelle Augenstein/Spandana Gella/Sebastian Ruder u.a. (Hg.), *Proc. of RepL4NLP-2019*, USA: ACL, S. 15–26.

Bühler, Stefan (5/17/2022), So bringt sich das Parlament an den Rand der Erschöpfung, *Aargauer Zeitung*, https://www.aargauerzeitung.ch/schweiz/flut-von-vorstoessen-nationalrat-erzielt-eigentor-wie-sich-das-parlament-selber-mit-arbeit-eindeckt-ld.2291431, 26.04.2023.

Bundesamt für Statistik (3/15/2022), Anteil der Wohnbevölkerung ab 16 Jahren mit einem grossen Vertrauen ins politische und Rechtssystem sowie in die Polizei, nach Migrationsstatus, verschiedenen soziodemografischen Merkmalen und Grossregion, https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/tabellen.assetdetail.21904835.html, 16.05.2023.

Bundesamt für Statistik (2023), Stimmbeteiligung, 16.05.2023, https://www.bfs.admin.ch/bfs/de/home/statistiken/politik/abstimmungen/stimmbeteiligung.html.

Bundesministerium für Inneres (2022), Aktionsplan Deepfake, 25.04.2023, https://bmi.gv.at/bmi_documents/2779.pdf.

Bundesrat (2017), Rechtliche Basis für Social Media: Erneute Standortbestimmung, 09.11.2023, https://www.bakom.admin.ch/dam/bakom/de/dokumente/informationsgesellschaft/social_media/social%20media%20bericht.pdf.download.pdf/social-media-bericht-2017-DE.pdf.

Bundesrat (2021), Ergänzungen betreffend Cybermobbing im Strafgesetzbuch: Bericht des Bundesrates, 04.11.2023, https://www.newsd.admin.ch/newsd/message/attachments/73646.pdf.

Bundesrat (Hg.) (2023a), Grosse Kommunikationsplattformen: Bundesrat strebt Regulierung an, 16.05.2023, https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-94116.html.

Bundesrat (2023b), Nationale Cyberstrategie (NCS), Bern.

Burchard, Hans von der (5/21/2018), Belgian socialist party circulates ‘deep fake’ Donald Trump video, *POLITICO*, https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/, 09.10.2022.

Burkart, Roland (2021), *Kommunikationswissenschaft*. *Grundlagen und Problemfelder einer interdisziplinären Sozialwissenschaft,* 6. vollst. überarb. u. aktual. Auflage, revidierte Ausgabe, Wien: UTB; Böhlau Verlag.

BuzzFeedVideo (2018), You Won’t Believe What Obama Says In This Video! :-), 09.10.2022, https://www.youtube.com/watch?v=cQ54GDm1eL0.

C2PA (2022), Coalition for Content Provenance and Authenticity, 13.05.2022, https://c2pa.org/.

Cahlan, Sarah (13.02.2020), How misinformation helped spark an attempted coup in Gabon, *The Washington Post*, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/, 16.11.2023.

Caplan, Robyn/danah boyd (2018), Isomorphism through algorithms: Institutional dependencies in the case of Facebook, *Big Data & Society*, Jg. 5, H. 1, 205395171875725.

Cappello, Maja (2020), *Künstliche Intelligenz im audiovisuellen Sektor*, Strasbourg: Europäische Audiovisuelle Informationsstelle.

Carlini, Nicholas/Hany Farid (2020), Evading Deepfake-Image Detectors with White- and Black-Box Attacks, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*: IEEE, S. 2804–2813.

Cavazos, Roberto (2019), The Economic Cost of Bad Actors on the Internet: Fake Influencer Marketing in 2019, 02.03.2023.

CFJM (2022), Fake news et deepfakes – Centre de formation au journalisme et aux médias, 05.05.2023, https://www.cfjm.ch/fake-news-et-deepfakes/.

Chan, Caroline/Shiry Ginosar/Tinghui Zhou/Alexei A. Efros (2019), Everybody Dance Now, 06.05.2022, https://carolineec.github.io/everybody_dance_now/.

Cheikosman, Evin/Nadia Hewett/Karin Gabriel (2021), Blockchain can help combat threat of deepfakes. Here's how, 17.11.2022, https://www.weforum.org/agenda/2021/10/how-blockchain-can-help-combat-threat-of-deepfakes/.

Chen, Chi-Ying/Zon-Ying Shae/Chien-Jen Chang/Kuan-Yuh Lin/Shu-Mei Tan/Shao-Liang Chang (2019), A Trusting News Ecosystem Against Fake News from Humanity and Technology Perspectives, in: Sanjay Misra (Hg.), *19th ICCSA*: IEEE, S. 132–137.

Chesney, Bobby/Danielle Citron (2019), Deep Fakes: A Looming Challenge for Privacy.

Chesney, Robert/Danielle Citron (2018), Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, *SSRN Electronic Journal*.

Chichizola, Corey (2020), *Rogue One Deepfake Makes Star Wars' Leia And Grand Moff Tarkin Look Even More Lifelike*, *CINEMABLEND*, https://www.cinemablend.com/news/2559935/rogue-one-deepfake-makes-star-wars-leia-and-grand-moff-tarkin-look-even-more-lifelike, 02.06.2022.

Christians, Clifford G./Theodore L. Glasser/Denis McQuail/Kaarle Nordenstreng/Robert A. White (2009), *Normative theories of the media*. *Journalism in democratic societies*, Urbana: University of Illinois Press.

Chu, Beilin/Weike You/Zhen Yang/Linna Zhou/Renying Wang (2022), Protecting World Leader Using Facial Speaking Pattern Against Deepfakes, *IEEE Signal Processing Letters*, Jg. 29, S. 2078–2082.

Ciftci, Umur Aybars/Ilke Demir/Lijun Yin (2020), FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals, *IEEE transactions on pattern analysis and machine intelligence*, PP.

Cirone, Alexandra/William Hobbs (2023), Asymmetric flooding as a tool for foreign influence on social media, *Political Science Research and Methods*, Jg. 11, H. 1, S. 160–171.

Citron, Danielle/Robert Chesney (2019), Deepfakes and the New Disinformation War, *Foreign Affairs*.

Cochran, Justin D./Stuart A. Napshin (2021), Deepfakes: Awareness, Concerns, and Platform Accountability, *Cyberpsychology, Behavior, and Social Networking*, Jg. 24, H. 3, S. 164–172.

Coldewey, Devin (2022), *This co-worker does not exist: FBI warns of deepfakes interviewing for tech jobs*, *TechCrunch*, https://techcrunch.com/2022/06/28/this-coworker-does-not-exist-fbi-warns-of-deepfakes-interviewing-for-tech-jobs/, 04.04.2023.

Collins, Aengus (2019), Forged Authenticity: Governing Deepfake Risks.

Collins, Aengus/Touradj Ebrahimi (2021), Risk governance and the rise of deepfakes, in: EPFL, 06.04.2022, https://www.epfl.ch/research/domains/irgc/spotlight-on-risk-series/risk-governance-and-the-rise-of-deepfakes/.

Compton, Sophie (2021), What Are Deepfakes? More and More Women Are Facing This Form of Online Abuse | Vogue, in: Vogue, https://www.vogue.com/article/scary-reality-of-deepfakes-online-abuse.

Corbu, Nicoleta/Denisa-Adriana Oprea/Elena Negrea-Busuioc/Loredana Radu (2020), 'They can't fool me, but they can fool the others!' Third person effect and fake news detection, *European Journal of Communication*, Jg. 35, H. 2, S. 165–180.

Council of Europe (2011), Übereinkommen des Europarats zur Verhütung und Bekämpfung von Gewalt gegen Frauen und häuslicher Gewalt und erläuternder Bericht. SR 0.311.35, *Council of Europe Treaty Series*, H. 210, S. 1–122.

Council of Europe (Hg.) (2016), Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom, 08.11.2023,

https://www.coe.int/en/web/freedom-expression/committee-of-ministers-adopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-cm-rec-2016-5-of-the-committee-of-ministers-to-member-states-on-internet-freedom.

Council on Foreign Relations (2023), Tracking State-Sponsored Cyberattacks Around the World, 16.11.2023, https://www.cfr.org/cyber-operations/.

Covello, Vincent T. (1983), The perception of technological risks: A literature review, *Technological Forecasting and Social Change*, Jg. 23, H. 4, S. 285–297.

Craig, Emory (1/15/2020), Samsung's Neon Project – Artificial Humans or Chatbots?, *Digital Bodies*, https://www.digitalbodies.net/samsungs-neon-project-artificial-humans-or-chatbots/, 09.10.2022.

Cueni, Raphaela, *Schutz von Satire im Rahmen der Meinungsfreiheit*. Dissertation.

Cueni, Raphaela (2019), Falsche und irreführende Informationen im Verfassungsrecht der Schweiz, *ex/ante*, H. 1, S. 3–17.

Dagar, Deepak/Dinesh Kumar Vishwakarma (2022), A literature review and perspectives in deepfakes: generation, detection, and applications, *International Journal of Multimedia Information Retrieval*, Jg. 11, H. 3, S. 219–289.

Daloz, Jocelyn (11/20/2021), Oskar Freysinger verbreitet Fake-Video von Berset, *Blick*, https://www.blick.ch/politik/umgeleitete-kampagne-oskar-freysinger-verbreitet-fake-video-von-berset-id16999942.html, 25.04.2023.

Damer, Naser/Alexandra Mosegui Saladie/Andreas Braun/Arjan Kuijper (2018), MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network, in: *9th BTAS*, NJ: IEEE, S. 1–10.

Das Schweizer Parlament (Hg.) (2021), Schutz der Schweizer Demokratie vor ausländischer Propaganda und Desinformation (2), 26.04.2023, https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20213385.

Debus, Marc/Hanna Bäck (2014), Der Einfluss von Mandatstyp und Wahlkreiseigenschaften auf die inhaltlichen Positionen in Parlamentsreden: Eine Untersuchung anhand der parlamentarischen Debatten im Deutschen Bundestag von 1998 bis 2002, *Swiss Political Science Review*, Jg. 20, H. 2, S. 330–353.

Deepbrain AI (2023), Create generative AI videos with just a script, 17.11.2023, https://www.deepbrain.io/.

Deepware (2023a), deepware repository, https://github.com/deepware.

Deepware (2023b), Frequently Asked Questions, https://deepware.ai/faq/.

Delnon, Vera/Bernhard Rüdy (2019), Kommentar zu Art. 180 StGB, in: Marcel Alexander Niggli/Hans Wiprächtiger/Jürg-Beat Ackermann (Hg.), *Strafrecht. Strafgesetzbuch, Jugendstrafgesetz*, 4. Auflage, Basel: Helbing Lichtenhahn Verlag.

Der Spiegel (02.07.2023), Bericht der Uno: Nordkoreas Hacker erbeuten Rekordsummen für Atomprogramm, *DER SPIEGEL*, https://www.spiegel.de/netzwelt/nordkorea-hacker-fuer-kim-jong-un-erbeuten-rekordsummen-fuer-atomprogramm-uno-bericht-a-8983b3e0-9504-42ed-a31b-5d08ab17c6e7, 17.05.2023.

Dessa (2019), Detecting Audio Deepfakes With AI, 13.09.2023, https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35.

Deutscher Bundestag (2020), Drucksache 19/23700.

Diakopoulos, Nicholas/Deborah Johnson (2021), Anticipating and addressing the ethical implications of deepfakes in the context of elections, *New Media & Society*, Jg. 23, H. 7, S. 2072–2098.

Diresta, Renee (7/31/2020), AI-Generated Text Is the Scariest Deepfake of All, *WIRED*, https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/, 16.11.2023.

Dobber, Tom/Nadia Metoui/Damian Trilling/Natali Helberger/Claes de Vreese (2021), Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?, *The International Journal of Press/Politics*, Jg. 26, H. 1, S. 69–91.

Dörner, Andreas/Benedikt Porzelt (2016), Politisches Gelächter. Rahmen, Rahmungen und Rollen bei Auftritten politischer Akteure in satirischen Interviews des deutschen Fernsehens, *Medien & Kommunikationswissenschaft*, Jg. 64, H. 3, S. 339–358.

Doublet, Yves-Marie (2019), Disinformation and electoral campaigns, https://rm.coe.int/disinformation-and-electoral-campaigns/16809fa91f.

Duffy, Andrew/Peng Hwa Ang (2019), Digital Journalism: Defined, Refined, or Re-defined, *Digital Journalism*, Jg. 7, H. 3, S. 378–385.

Dumitru, Elena-Alexandra (2021), Is “Letting the Truth Get in the Way of a Good Story” Enough? Journalists’ Perception on Fake News, *Journal of Media Research*, Jg. 14, 3 (41), S. 63–79.

EC (Hg.) (2022a), Signatories of the 2022 Strengthened Code of Practice on Disinformation, 09.11.2023, https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation.

EC (Hg.) (2022b), The 2022 Code of Practice on Disinformation, 09.11.2023, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

EDSB (2018), Stellungnahme des EDSB zu Online-Manipulation und personenbezogenen Daten. Stellungnahme 3/2018.

Edwards, Benj (2023), Why AI detectors think the US Constitution was written by AI, 18.08.2023, https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/.

EFD (2023), Faktencheck, 16.05.2023, https://www.efd.admin.ch/efd/de/home/das-efd/nsb-news_list/faktencheck.html.

ENISA (2023), Identifying emerging cybersecurity threats and challenges for 2030, Luxembourg, 20.04.2023, https://data.europa.eu/doi/10.2824/117542.

Enste, Dominik H/Lena Suling (2020), Vertrauen in Wirtschaft, Staat, Gesellschaft 2020, *Enste IW-Policy Paper*, H. 5.

Epstein-Gross, Casey (26.07.2023), OpenAI Abruptly Shuts Down ChatGPT Plagiarism Detector – And Educators Are Worried, *Observer*, https://observer.com/2023/07/openai-shut-ai-classifier/, 18.08.2023.

EPTA (2023), Generative Artificial Intelligence. Opportunities, Risks And Policy Challenges, https://www.parlament.cat/document/composicio/394503200.pdf.

Espeloer, Mathias (2022), IT-Equipment und Data Security für mobiles Arbeiten, S. 187–204.

Esser, Patrick/Johnathan Chiu/Parmida Atighehchian/Jonathan Granskog/Anastasis Germanidis (2023), Structure and Content-Guided Video Synthesis with Diffusion Models, https://arxiv.org/abs/2302.03011.

Ettema, Yori (2021), Deepmemory, 17.05.2022, https://www.yorie.nl/deepmemory/.

euronews (2023), EU-Polizeibehörde warnt vor Missbrauch von KI durch Kriminelle, 22.09.2023, https://de.euronews.com/next/2023/05/09/europol-ist-besorgt-uber-die-moglichkeiten-von-chatgpt-fur-verbrechen-hier-die-grunde.

Europäische Kommission (Hg.) (2021a), Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the transparency and targeting of political advertising. COM(2021) 731 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0731.

Europäische Kommission (2021b), Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, S. 1–120.

European Commission (2018), A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation., https://data.europa.eu/doi/10.2759/739290.

European Commission (2022), Open-source intelligence. The lessons that OSINT provides to open-data portals, in: European Commission, 27.06.2023, https://data.europa.eu/en/publications/datastories/open-source-intelligence.

Europol (2020), Malicious Uses and Abuses of Artificial Intelligence, 07.12.2020, https://thelivinglib.org/malicious-uses-and-abuses-of-artificial-intelligence/.

Europol (2022), Facing reality?, 18.05.2022, https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

Fallis, Don (2021), The Epistemic Threat of Deepfakes, *Philosophy & technology*, Jg. 34, H. 4, S. 623–643.

Farid, Hany/Hans-Jakob Schindler (2020), *Deepfakes. Eine Bedrohung für Demokratie und Gesellschaft*, Berlin: Konrad-Adenauer-Stiftung e.V.

FBI (2021), Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations. Private Industry Notification.

FBI (2022), *Internet Crime Complaint Center (IC3) Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions*, https://www.ic3.gov/Media/Y2022/PSA220628, 22.07.2022.

Feeney, Matthew (2021), Deepfake Laws Risk Creating More Problems Than They Solve, https://rtp.fedsoc.org/wp-content/uploads/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf.

Fellmann, Walter (2011), Kommentierung zu Art. 12 BFA, in: Walter Fellmann/Gaudenz G. Zindel (Hg.), *Kommentar zum Anwaltsgesetz*, 2. Auflage, Genf/Zürich/Basel: SwissLex; Schulthess Juristische Medien AG.

Fernando, Tharindu/Clinton Fookes/Simon Denman/Sridha Sridharan (2019), Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks, https://arxiv.org/pdf/1911.07844.

Fiolka, Gerhard (2019), Kommentierung zu Art. 258 StGB, in: Marcel Alexander Niggli/Hans Wiprächtiger/Jürg-Beat Ackermann (Hg.), *Strafrecht. Strafgesetzbuch, Jugendstrafgesetz*, 4. Auflage, Basel: Helbing Lichtenhahn Verlag.

Foley, Joseph (2023), *20 of the best deepfake examples that terrified and amused the internet*, *Creative Bloq*, https://www.creativebloq.com/features/deepfake-examples, 27.06.2023.

Forbes (2021), The Rise Of Voice Cloning And DeepFakes In The Disinformation Wars, 06.05.2022, https://www.forbes.com/sites/jenniferhicks/2021/09/21/the-rise-of-voice-cloning-and-deep-fakes-in-the-disinformation-wars/.

Foster, Peter (23.04.2013), "Bogus" AP tweet about explosion at the White House wipes billions off US markets, *The Telegraph*, https://www.telegraph.co.uk/finance/markets/10013768/Bogus-AP-tweet-about-explosion-at-the-White-House-wipes-billions-off-US-markets.html, 26.04.2023.

Fraga-Lamas, Paula/Tiago M. Fernandez-Carames (2020), Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality, *IT Professional*, Jg. 22, H. 2, S. 53–59.

Frewer, Lynn J./Chaya Howard/Duncan Hedderley/Richard Shepherd (1999), Reactions to information about genetic engineering: impact of source characteristics, perceived personal relevance, and persuasiveness, *Public Understanding of Science*, Jg. 8, H. 1, S. 35.

Fulterer, Ruth (10/24/2022), Deepfake-Anruf beim Hersteller der Bayraktar-Drohne sollte die Ukraine diskreditieren, *Neue Zürcher Zeitung*, https://www.nzz.ch/technologie/deepfake-anruf-beim-hersteller-der-bayraktar-drohe-ld.1708494, 25.04.2023.

Gafni, Oran/Oron Ashual/Lior Wolf (2021), Single-Shot Freestyle Dance Reenactment, in: *Proc. of 2021 IEEE/CVF*, Piscataway, NJ: IEEE, S. 882–891.

Gamage, Dilrukshi/Jiayu Chen/Piyush Ghasiya/Kazutoshi Sasahara (2022a), Deepfakes and Society: What Lies Ahead?, in: Mahdi Khosravy/Isao Echizen/Noboru Babaguchi (Hg.), *Frontiers in Fake Media Generation and Detection*, 1st ed. 2022, Singapore: Springer, S. 3–43.

Gamage, Dilrukshi/Piyush Ghasiya/Vamshi Krishna Bonagiri/Mark E. Whiting/Kazutoshi Sasahara (2022b), Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications, in: *CHI Conference on Human Factors in Computing Systems*.

Gentzel, Peter (2022), Andreas Hepp (2021). Auf dem Weg zur digitalen Gesellschaft. Über die tiefgreifende Mediatisierung der sozialen Welt, *Publizistik*, Jg. 67, 2-3, S. 397–400.

Gewirtz, David (08.11.2023), How to write better ChatGPT prompts for the best generative AI results, *ZDNET*, https://www.zdnet.com/article/how-to-write-better-chatgpt-prompts/, 22.09.2023.

GI (2022), Deep Learning, 15.11.2022, https://gi.de/informatiklexikon/deep-learning.

Gjon, David (2019), 22 plädoyer 4/19, Wahrheitsfindung mit Virtual-Reality-Brille, *plädoyer*, H. 4, S. 22–25.

Godulla, Alexander/Christian P. Hoffmann/Daniel Seibert (2021), Dealing with deepfakes – an interdisciplinary examination of the state of research and implications for communication studies, *Studies in Communication and Media*, Jg. 10, H. 1, S. 72–96.

Goodin, Dan (2023), *Hackers are selling a service that bypasses ChatGPT restrictions on malware*, *Ars Technica*, https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/, 14.04.2023.

Google Developers (2020), Generative Adversarial Networks. Common Problems, 12.05.2022, https://developers.google.com/machine-learning/gan/problems.

Google Developers (2022), Häufige Probleme | Machine Learning, 14.11.2022, https://developers.google.com/machine-learning/gan/problems.

Google Developers (2023), Imagen Video, in: https://imagen.research.google/video, 17.11.2023, https://imagen.research.google/video/.

Google Inc. (Hg.) (2023), Richtlinien zu Fehlinformationen, 09.11.2023, https://support.google.com/youtube/answer/10834785?hl=de.

Gosse, Chandell/Jacquelyn Burkell (2020), Politics and porn: how news media characterizes problems presented by deepfakes, *Critical Studies in Media Communication*, Jg. 37, H. 5, S. 497–511.

GPT-3 (08.09.2020), A robot wrote this entire article. Are you scared yet, human?, *The Guardian*, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3, 09.05.2022.

Gregory, Sam (2019), Preparing for deepfakes against journalism, *NEWS REPORT 2019*, S. 105–107.

GREVIO (2021), GREVIO publishes its General Recommendation No.1 on the digital dimension of violence against women, 09.11.2023, https://www.coe.int/en/web/istanbul-convention/-/grevio-publishes-its-general-recommendation-no-1.

Grinberg, Nir/Kenneth Joseph/Lisa Friedland/Briony Swire-Thompson/David Lazer (2019), Fake news on Twitter during the 2016 U.S. presidential election, *Science (New York, N.Y.)*, Jg. 363, H. 6425, S. 374–378.

Groh, Matt (n.n.), Project Overview Detect DeepFakes: How to counteract misinformation created by AI, 25.04.2023, https://www.media.mit.edu/projects/detect-fakes/overview/.

Groner, Roger (2011), *Beweisrecht*, Genf/Zürich/Basel/Bern: SwissLex; Stämpfli Verlag AG.

Guess, Andrew/Jonathan Nagler/Joshua Tucker (2019), Less than you think: Prevalence and predictors of fake news dissemination on Facebook, *Science advances*, Jg. 5, H. 1, eaau4586.

Guess, Andrew M./Michael Lerner/Benjamin Lyons/Jacob M. Montgomery/Brendan Nyhan/Jason Reifler/Neelanjan Sircar (2020), A digital media literacy intervention increases discernment between mainstream and false news in the United States and India, *Proceedings of the National Academy of Sciences*, Jg. 117, H. 27, S. 15536–15545.

Guglielmi, Giorgia (2020), The next-generation bots interfering with the US election, *Nature*, Jg. 587, H. 7832, S. 21.

Gupta, Sneha (2020), Disadvantages of GANs. Am I real or a Trained Model to write?, 10.11.2022, https://iq.opengenus.org/disadvantages-of-gans/.

Gustafsod, Per E. (1998), Gender Differences in Risk Perception: Theoretical and Methodological erspectives, *Risk Analysis*, Jg. 18, H. 6, S. 805–811.

Gutsche, Robert E. (2019), The State and Future of Television News Studies: Theoretical Perspectives, Methodological Problems, and Practice, *Journalism Practice*, Jg. 13, H. 9, S. 1034–1041.

Habermas, Jürgen (2022), *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik,* Erste Auflage, Originalausgabe, Berlin: Suhrkamp.

Hameleers, Michael/Toni G. L. A. van der Meer/Tom Dobber (2022), You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media, *Social Media + Society*, Jg. 8, H. 3, 205630512211163.

Handan-Nader, Cassandra (2023), Do fake online comments pose a threat to regulatory policymaking? Evidence from Internet regulation in the United States, *Policy & Internet*, Jg. 15, H. 1, S. 100–116.

Hansjakob, Thomas (2018), *Überwachungsrecht der Schweiz*. *Kommentar zu Art. 269 ff. StPO und zum BÜPF,* [1. Auflage], Zürich/Basel/Genf: Schulthess.

Hao, Karen (4/13/2021), Deepfake-Tweets trollen Amazon, *heise online*, https://www.heise.de/hintergrund/Deepfake-Tweets-trollen-Amazon-6011681.html, 16.05.2023.

Hargittai, Eszter/Yuli Patrick Hsieh (2012), Succinct Survey Measures of Web-Use Skills, *Social Science Computer Review*, Jg. 30, H. 1, S. 95–107.

Hargittai, Eszter/Anne Marie Piper/Meredith Ringel Morris (2019), From internet access to internet skills: digital inequality among older adults, *Universal Access in the Information Society*, Jg. 18, H. 4, S. 881–890.

Hart, Jordan/Aaron Mok (2023), College professors are going back to paper exams and handwritten essays to fight students using ChatGPT, in: businessinsider, 18.08.2023, https://www.businessinsider.com/chatgpt-driving-return-to-paper-exams-written-essays-at-universities-2023-8.

Harwell, Drew (24.05.2019), Faked Pelosi videos, slowed to make her appear drunk, spread across social media, *The Washington Post*, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/.

Hasenböhler, Franz (2019), *Das Beweisrecht der ZPO*, Zürich/Basel/Genf: Schulthess.

Hausheer, Heinz/Regina E. Aebi-Müller (2020), *Das Personenrecht des Schweizerischen Zivilgesetzbuches,* 5. Auflage, Bern: Stämpfli Verlag.

Heesen, Jessica/Christoph Bieber/Armin Grunwald/Tobias Matzner/Alexander Roßnagel (2021), KI-Systeme und die individuelle Wahlentscheidung.

Heinemann, Andreas/Beat Althaus (2015), Posten, Liken, Sharen – Urheberrecht in sozialen Netzwerken, *jusletter*, H. 12.

Helberger, Natali (2019), On the Democratic Role of News Recommenders, *Digital Journalism*, Jg. 7, H. 8, S. 993–1012.

Helfenberger, Leo (07.04.2023), «Irreführend»: Darum sorgt die FDP mit einem KI-Bild zu Klimaklebern gerade für rote Köpfe, *watson*, https://www.watson.ch/schweiz/wahlen%202023/199284080-fdp-macht-wahlkampf-mit-ki-bild-warum-nun-darueber-diskutiert-wird, 16.11.2023.

Hellyer, Francis (22.03.2022), Deepfakes: The New Ticket to Immortality?, *RollingStone*, https://www.rollingstone.com/culture-council/articles/the-new-ticket-to-immortality-1324513/.

Hertig, Maja (2015a), Kommentierung zu Art. 16 BV, in: Bernhard Waldmann/Eva Maria Belser/Astrid Epiney (Hg.), *Bundesverfassung*, Basel: Helbing-Lichtenhahn-Verl.

Hertig, Maja (2015b), Vorbemerkungen zu den Kommunikationsgrundrechten, in: Bernhard Waldmann/Eva Maria Belser/Astrid Epiney (Hg.), *Bundesverfassung*, Basel: Helbing-Lichtenhahn-Verl.

Hewage, Chaminda (2020), Data Protection in the Wake of Deepfakes, 04.11.2023, https://www.infosecurity-magazine.com/next-gen-infosec/data-protection-wake-deepfakes/.

High Level Expert Group (HLEG) (2018), A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation., https://www.doi.org/10.2759/739290.

Hiort, Astrid (2022), The Most-Followed Virtual Influencers of 2022, 26.04.2023, https://www.virtualhumans.org/article/the-most-followed-virtual-influencers-of-2022.

Ho, Jonathan/William Chan/Chitwan Saharia/Jay Whang/Ruiqi Gao/Alexey Gritsenko/Diederik P. Kingma/Ben Poole/Mohammad Norouzi/David J. Fleet/Tim Salimans (2022), Imagen Video: High Definition Video Generation with Diffusion Models.

Hochstrasser, Judith (2. März 2023), «Ich finde es krass, wenn mit Fakes politisiert wird», *Horizonte – Das Schweizer Forschungsmagazin*, https://www.horizonte-magazin.ch/2023/03/02/ich-finde-es-krass-wenn-mit-fakes-politisiert-wird/, 27.06.2023.

Hong, Yangsun/Sunghak Kim (2020), Influence of Presumed Media Influence for Health Prevention: How Mass Media Indirectly Promote Health Prevention Behaviors through Descriptive Norms, *Health communication*, Jg. 35, H. 14, S. 1800–1810.

Hsu, Chih-Chung/Yi-Xiu Zhuang/Chia-Yen Lee (2020), Deep Fake Image Detection Based on Pairwise Learning, *Applied Sciences*, Jg. 10, H. 1, S. 370.

Huang, Kung-Hsiang/Kathleen McKeown/Preslav Nakov/Yejin Choi/Heng Ji (2022), Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation.

Humprecht, Edda/Frank Esser/Peter van Aelst/Anna Staender/Sophie Morosoli (2021), The sharing of disinformation in cross-national comparison: analyzing patterns of resilience, *Information, Communication & Society*, Jg. 26, H. 7, S. 1–21.

Hwang, Yoori/Ji Youn Ryu/Se-Hoon Jeong (2021), Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education, *Cyberpsychology, Behavior, and Social Networking*, Jg. 24, H. 3, S. 188–193.

IMF (2022), *Download entire World Economic Outlook database, October 2022*, *IMF*, https://www.imf.org/en/Publications/WEO/weo-database/2022/October/download-entire-database, 20.04.2023.

Insikt Group (2023), The Business of Fraud: Deepfakes, Fraud's Next Frontier, 26.04.2023, https://go.recordedfuture.com/hubfs/reports/cta-2021-0429.pdf.

ITU (2021), Global Cybersecurity Index 2020, Geneva, Switzerland.

Jacquemin, Quentin (2023), Le droit suisse permet-il de réprimer les deepfakes?, in: Florence Guillaume (Hg.), *La technologie, l'humain et le droit*.

Jahng, Mi Rosie/Stine Eckert/Jade Metzger-Riftkin (2023), Defending the Profession: U.S. Journalists' Role Understanding in the Era of Fake News, *Journalism Practice*, Jg. 17, H. 2, S. 226–244.

Jalalifar, Seyed A./Hosein Hasani/Hamid Aghajan (2018), Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks, https://arxiv.org/pdf/1803.07461.

JAM (2019a), Comment Reuters forme ses journalistes à repérer les "Deepfakes", 05.05.2023, https://jam.unine.ch/comment-reuters-forme-ses-journalistes-a-reperer-les-deepfakes/.

JAM (2019b), Les "Deepfakes", le nouveau fléau d'internet, 05.05.2023, https://jam.unine.ch/les-deepfakes-le-nouveau-fleau-dinternet/.

Jarke, Juliane (2018), Digitalisierung und Gesellschaft, *Soziologische Revue*, Jg. 41, H. 1, S. 3–20.

Jarrahi, Ali/Leila Safari (2022), Evaluating the effectiveness of publishers' features in fake news detection on social media, *Multimedia tools and applications*, S. 1–27.

Jarren, Otfried/Renate Fischer (2022), Transformation der politischen Öffentlichkeit? Der Einfluss von Plattformen auf das gesellschaftliche Vermittlungssystem, *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Jg. 74, S1, S. 183–207.

Jaun, René/Yannick Züllig (2023), Update: Meldepflicht für Cyberangriffe kommt – jene für Schwachstellen nicht, 16.11.2023, https://www.swisscybersecurity.net/cybersecurity/2022-05-13/meldepflicht-fuer-cyberangriffe-auf-kritische-infrastrukturen-stoesst-auf.

Jin, Youngjin/Eugene Jang/Jian Cui/Jin-Woo Chung/Yongjae Lee/Seungwon Shin (2023), DarkBERT: A Language Model for the Dark Side of the Internet, http://arxiv.org/pdf/2305.08596v2.

Jugend und Medien (Hg.) (2023), Fake News & Manipulation – von Bubbles, Bots und Hoaxes, 08.11.2023, https://www.jugendundmedien.ch/themen/fake-news-manipulation.

Juneja, Prerna/Tanushree Mitra (2022), Human and technological infrastructures of fact-checking.

Junghärtchen, Immo (2023), Videoproduktion mit künstlicher Intelligenz umsetzen, in: Heise Medien, 17.11.2023, https://www.heise.de/ratgeber/Videoproduktion-mit-kuenstlicher-Intelligenz-umsetzen-7544009.html?seite=all.

Jungherr, Andreas (2022), Digital campaigning: How digital media change the work of parties and campaign organizations and impact elections.

Jungherr, Andreas/Ralph Schroeder (2023), Artificial intelligence and the public arena, *Communication Theory*, Jg. 33, 2-3, S. 164–173.

Kägi, Willi Irene (2022), Deepfakes – echter Mehrwert oder unlautere Manipulation?, 06.04.2022, https://www.kalaidos-fh.ch/de-CH/Blog/Posts/2021/07/Digitaliserung-1040-Deepfakes-Mehrwert-oder-Manipulation.

Kaiber (2023), Discover the artist within you, 17.11.2023, https://kaiber.ai/.

Kalpokas, Ignas/Julija Kalpokiene (2022), Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation (SpringerBriefs in Political Science).

Kalsnes, Bente/Kajsa Falasca/Aske Kammer (2021), Scandinavian political journalism in a time of fake news and disinformation.

Karaboga, Murat (2023), Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags, in: Sylvia Jaki/Stefan Steiger (Hg.), *Digitale Hate Speech. Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation*: Springer, S. 197–220.

Keller, Erin (3/28/2023), Pope Francis in Balenciaga deepfake fools millions: 'Definitely scary', *New York Post*, https://nypost.com/2023/03/27/pope-francis-in-balenciaga-deepfake-fools-millions-definitely-scary/, 02.05.2023.

Kersten, Jens (2020), KI als Künstlerin, CAS LMU Blog.

Kim, Hyeongwoo/Pablo Garrido/Ayush Tewari/Weipeng Xu/Justus Thies/Matthias Niessner/Patrick Pérez/Christian Richardt/Michael Zollhöfer/Christian Theobalt (2018), Deep video portraits, *ACM Transactions on Graphics*, Jg. 37, H. 4, S. 1–14.

Kirchengast, Tyrone (2020), Deepfakes and image manipulation: criminalisation and control, *Information & Communications Technology Law*, Jg. 29, H. 3, S. 308–323.

Kleine, Felix (2022), *Perception of Deepfake Technology – The Influence of the Recipients' Affinity for Technology on the Perception of Deepfakes*.

Knorre et al. (2020), *Big-Data-Debatte*: Springer Nature.

Köbis, Nils/Barbora Doležalová/Ivan Soraperra (2021), Fooled Twice – People Cannot Detect Deepfakes But Think They Can, *SSRN Electronic Journal*.

Koc, Mustafa/Esra Barut (2016), Development and validation of New Media Literacy Scale (NMLS) for university students, *Computers in Human Behavior*, Jg. 63, S. 834–843.

Koch, Marie-Claire (2023), Nvidia erweitert Stable Diffusion zum hochauflösenden Text-zu-Video-Generator, 17.11.2023, https://www.heise.de/news/Video-LDM-Hochaufloesender-Text-zu-Video-Generator-aus-Nvidias-AI-Lab-8975541.html.

Koliska, Michael/Karin Assmann (2021), Lügenpresse: The lying press and German journalists' responses to a stigma, *Journalism*, Jg. 22, H. 11, S. 2729–2746.

König, Pascal D./Georg Wenzelburger (2020), Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics, *Government Information Quarterly*, Jg. 37, H. 3, S. 101489.

Koopman, Marissa/Andrea Macarulla Rodriguez/Zeno Geradts (2018), *Detection of Deepfake Video Manipulation*.

Korshunov, Pavel/Sebastien Marcel (2018), Speaker Inconsistency Detection in Tampered Video, in: *26th EUSIPCO*: IEEE, S. 2375–2379.

Kötke, Jennifer-Tia (2021), Deepfake - Eine kurze Einleitung, *Fakultät für Mathematik, Informatik und Naturwissenschaften, Universität Hamburg*.

Krebs, Dagmar/Natalja Menold (2019), Gütekriterien quantitativer Sozialforschung, in: Nina Baur/Jörg Blasius (Hg.), *Handbuch Methoden der empirischen Sozialforschung*, Wiesbaden: Springer Fachmedien Wiesbaden, S. 489–504.

Krohn, Jon/Grant Beyleveld/Aglaé Bassens (2020), *Deep Learning illustriert. Eine anschauliche Einführung in Machine Vision, Natural Language Processing und Bilderzeugung für Programmierer und Datenanalysten*, Heidelberg: dpunkt.verlag.

Kuhlmann, Nico (2020), Deepfakes: Was tun gegen gefälschte Videos?, 04.11.2023, https://www.lto.de/recht/hintergruende/h/deepfakes-gegen-gefaelschte-videos-vorgehen-gericht-persoenlichkeitsrecht-strafrecht/.

Kuhn, Johannes (12.06.2019), Zuckerbergs böser Zwilling bleibt auf Instagram, *Süddeutsche Zeitung*, https://www.sueddeutsche.de/digital/deepfake-facebook-zuckerberg-instagram-1.4483936, 26.04.2023.

Kuzniar, Nadia (2017), *Inpflichtnahme der Internet-Provider bei Urheberrechtsverletzungen: Ist die Umsetzung der Providerhaftung im Vorentwurf vom 15. Dezember 2015 zum URG geglückt?*: Schulthess.

Lamar, Kendrick (2022), The Heart Part 5, 09.10.2022, https://www.youtube.com/watch?v=uAPUkgeiFVY.

Lampert, Calvin (2023), Kryptobetrüger nutzen einen Elon-Musk-Deepfake, 16.05.2023, https://www.swisscybersecurity.net/cybersecurity/2023-05-03/kryptobetrueger-nutzen-einen-elon-musk-deepfake.

Langguth, Johannes/Konstantin Pogorelov/Stefan Brenner/Petra Filkuková/Daniel Thilo Schroeder (2021), Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes, *Frontiers in Communication*, Jg. 6.

Langhart, Nicole (2021), Deepfakes: Wie die AI-Technologie eindrückliche Marketing Massnahmen ermöglicht, 17.05.2022, https://marketing.ch/deepfakes-wie-die-ai-technologie-eindrueckliche-marketing-massnahmen-ermoeglicht/.

Lantwin, Tobias (2019), Deep Fakes – Düstere Zeiten für den Persönlichkeitsschutz?, *MMR*, H. 9, S. 557–632.

Laufenburg, Robin (2021), Wenn KI Gesichter und Stimmen generiert, 22.09.2023, https://www.pcspezialist.de/blog/2021/04/06/deepfake/.

Lazer, David M. J./Matthew A. Baum/Yochai Benkler/Adam J. Berinsky/Kelly M. Greenhill/Filippo Menczer/Miriam J. Metzger/Brendan Nyhan/Gordon Pennycook/David Rothschild/Michael Schudson/Steven A. Sloman/Cass R. Sunstein/Emily A. Thorson/Duncan J. Watts/Jonathan L. Zittrain (2018), The science of fake news, *Science (New York, N.Y.)*, Jg. 359, H. 6380, S. 1094–1096.

Lecheler, Sophie/Jana Laura Egelhofer (2022), Disinformation, Misinformation, and Fake News, in: Jesper Strömbäck/Åsa Wikforss (Hg.), *Knowledge resistance in high-choice information environments*, 1st Edition, New York NY: Routledge, S. 69–87.

Lee, Jiyoung/Soo Yun Shin (2022), Something that They Never Said: Multimodal Disinformation and Source Vividness in Understanding the Power of AI-Enabled Deepfake News, *Media Psychology*, Jg. 25, H. 4, S. 531–546.

Leetaru, Kalev (2018), Why Digital Signatures Won't Prevent Deep Fakes But Will Help Repressive Governments, 17.11.2022, https://www.forbes.com/sites/kalevleetaru/2018/09/09/why-digital-signatures-wont-prevent-deep-fakes-but-will-help-repressive-governments/.

Lehrplan 21 (2016), Die Schülerinnen und Schüler können Medien und Medienbeiträge entschlüsseln, reflektieren und nutzen, 11.12.2023, https://v-fe.lehrplan.ch/index.php?code=a%7C10%7C0%7C1%7C0%7C2.

Leibowicz, Claire/Jonathan Stray/Emily Saltz (2020, July 13), Manipulated Media Detection Requires More Than Tools: Community Insights on Whats Needed - The Partnership on AI, in: Partnership on AI, https://www.partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/.

Leukfeldt, Eric Rutger/Majid Yar (2016), Applying Routine Activity Theory to Cybercrime: A Theoretical and Empirical Analysis, *Deviant Behavior*, Jg. 37, H. 3, S. 263–280.

Lewis, Andrew/Patrick Vu/Raymond Duch/Areeq Chowdhury (2022), Do Content Warnings Help People Spot a Deepfake? Evidence from Two Experiments.

Lewis, Jake (2023), “First time I’ve seen this”, 16.05.2023, https://twitter.com/jake____lewis/status/1630036971388108806.

Lewke, Christian (2017), “…aber das kann ich nicht tun!”: Künstliche Intelligenz und ihre Beteiligung am öffentlichen Diskurs, *InTeR*.

Li, Lingzhi/Jianmin Bao/Hao Yang/Dong Chen/Fang Wen (2019), FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping, http://arxiv.org/pdf/1912.13457v3.

Li, Lingzhi/Jianmin Bao/Ting Zhang/Hao Yang/Dong Chen/Fang Wen/Baining Guo (2020), Face X-Ray for More General Face Forgery Detection, in: Eric Mortensen/Margaux Masson-Forsythe (Hg.), *Proc. of IEEE/CVF*: IEEE, S. 5000–5009.

Li, Yuezun/Siwei Lyu (2019), Exposing DeepFake Videos By Detecting Face Warping Artifacts, 12.05.2022, https://github.com/yuezunli/DSP-FWA.

Lima, Oscar de/Sean Franklin/Shreshtha Basu/Blake Karwoski/Annet George (2020), Deepfake Detection using Spatiotemporal Convolutional Networks, https://arxiv.org/pdf/2006.14749.

Lin, Bibo/Seth C. Lewis (2022), The One Thing Journalistic AI Just Might Do for Democracy, *Digital Journalism*, Jg. 10, H. 10, S. 1627–1649.

Linardato, Dimitrios (2021), Deepfakes regulieren: Das Ende der Verlässlichkeit?, in: Lto, 06.11.2023, https://www.lto.de/recht/hintergruende/h/deepfakes-regulierung-europa-eu-schaden-demokratie-manipulation/.

Ling, Zhen-Hua/Xiao Zhou/Simon King (2021), The Blizzard Challenge 2021, 02.08.2022, http://festvox.org/blizzard/bc2021/BC21_ling_zhou_king.pdf.

Lu, Hang/Haoran Chu (2023), Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts, *Computers in Human Behavior*, Jg. 145, S. 107761.

Ludewig, Claus (03.08.2023), Vorsicht vor neuem Enkeltrick: Betrüger arbeiten mit KI-Tool, *PC Games Hardware*, https://www.pcgameshardware.de/Neue-Technologien-Thema-71240/News/Neuer-Enkeltrick-mit-Deepfake-1414907/, 16.05.2023.

Lyu, Siwei (2020), Deepfake Detection: Current Challenges and Next Steps, in: *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, S. 1–6.

Mäder, Lukas (2023), Cyberangriff auf SBB: Die Analyse läuft noch, *Neue Zürcher Zeitung*.

Mäder, Stefan/Marcel Alexander Niggli (2019), Kommentierung zu Art. 146 StGB, in: Marcel Alexander Niggli/Hans Wiprächtiger/Jürg-Beat Ackermann (Hg.), *Strafrecht. Strafgesetzbuch, Jugendstrafgesetz*, 4. Auflage, Basel: Helbing Lichtenhahn Verlag.

Magnani, Marco (2022), Microtargeting, Divisive Campaigns and the Rise in Voter Polarization, *SSRN Electronic Journal*.

Mahbub, Syed/Eric Pardede/A. S. M. Kayes/Wenny Rahayu (2019), Controlling astroturfing on the internet: a survey on detection techniques and research challenges, Int. J. Web and Grid Services, Vol. 15, No. 2, 2019 139.

Maksl, Adam/Seth Ashley/Stephanie Craft (2015), Measuring News Media Literacy, *Journal of Media Literacy Education*, Jg. 6, H. 3, S. 29–45.

Malik, Asad/Minoru Kuribayashi/Sani M. Abdullahi/Ahmad Neyaz Khan (2022), DeepFake Detection for Human Face Images and Videos: A Survey, *IEEE Access*, Jg. 10, S. 18757–18775.

Marchetti, Sabina (2022), Rolling in the Deep(Fakes). SSRN Scholarly Paper, Rochester, NY, 21.04.2022.

Marconi, F./T. Daldrup (2018), How The Wall Street Journal is preparing its journalists to detect deepfakes, https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes.

Marti, Arnold (2000), Selbstregulierung anstelle staatlicher Gesetzgebung?, *ZBL*, H. 101.

Martin, Kim (2020), What is Voice Cloning?, 06.05.2022, https://www.idrnd.ai/what-is-voice-cloning/.

Mason, Jennifer (2018), *Qualitative researching,* 3rd edition, Thousand Oaks CA: SAGE Publications.

Masood, Momina/Mariam Nawaz/Khalid Mahmood Malik/Ali Javed/Aun Irtaza/Hafiz Malik (2022), Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward, *Applied Intelligence*.

Mayring, Philipp (2010), *Qualitative Inhaltsanalyse*. *Grundlagen und Techniken,* 11. Neuausgabe, Weinheim: Beltz.

MAZ (2023), Links, Tools und Literatur Journalismus und Medien, 05.05.2023, https://www.maz.ch/publikation/links-tools-und-literatur-zu-journalismus-und-medien.

McChesney, Robert W. (2008), *The political economy of media*. *Enduring issues, emerging dilemmas / Robert W. McChesney*, New York: Monthly Review.

McCosker, Anthony (2022), Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube, *New Media & Society*, 146144482210939.

Meckel, Miriam/Léa Steinacker (2021), Hybrid Reality: The Rise of Deepfakes and Diverging Truths, *Morals & Machines*, Jg. 1, H. 1, S. 10–21.

Meinicke, Dirk (2020), Strafrechtliche Aspekte sogenannter Deep Fakes, *Den @ Wandel begleiten; IT-rechtliche Herausforderungen der Digitalisierung*.

Meiritz, Annett/Dana Heide/Christoph Herwartz (7/19/2021), Five Eyes: Internationale Allianz gegen chinesische Cyberangriffe, *Handelsblatt*, https://www.handelsblatt.com/politik/international/hackerangriffe-gegen-chinas-boeswillige-cyberaktivitaeten-usa-schmieden-allianz-mit-europa-und-japan/27433112.html, 17.05.2023.

Melli, Andreas (2022), Kommentierung zu Art. 28 ZGB, in: Thomas Geiser/Christiana Fountoulakis (Hg.), *Zivilgesetzbuch I. Art. 1-456 ZGB*, 7. Auflage: Basler Kommentar.

Menkens, Sabine (11/21/2019), Darknet: Um in die Foren zu kommen, müssen Pädophile Kinderpornos liefern, *WELT*, https://www.welt.de/politik/deutschland/article203673432/Darknet-Um-in-die-Foren-zu-kommen-muessen-Paedophile-Kinderpornos-liefern.html, 08.11.2023.

Meta (Hg.) (2022), Introducing Make-A-Video: An AI system that generates videos from text, 16.05.2023, https://ai.facebook.com/blog/generative-ai-text-to-video/.

Meta (Hg.) (2023), Misinformation, 09.11.2023, https://transparency.fb.com/de-de/policies/community-standards/misinformation/.

Meta AI (2023), Make-A-Video, 17.11.2023, https://makeavideo.studio/.

Metaphysic.ai (2022), The Future of Generative Adversarial Networks in Deepfakes – Metaphysic.ai, 16.11.2022, https://metaphysic.ai/the-future-of-generative-adversarial-networks-in-deepfakes/.

Metzger, Nils/Jan Schneider (2022), Wie Deepfakes im Ukraine-Krieg genutzt werden, 11.05.2022, https://www.zdf.de/nachrichten/politik/selenskyj-deepfake-video-ukraine-krieg-russland-100.html.

Meuser, Michael/Ulrike Nagel (1991), ExpertInneninterviews – vielfach erprobt, wenig bedacht, in: Detlef Garz/Klaus Kraimer (Hg.), *Qualitativ-empirische Sozialforschung*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 441–471.

Microsoft Innovation (Hg.) (2022), Explore Project Origin, 13.05.2022, https://innovation.microsoft.com/en-us/exploring-project-origin.

Mirchandani, Maya (2020), *Tackling Insurgent Ideologies in a Pandemic World*.

Mirsky, Yisroel/Wenke Lee (2022), The Creation and Detection of Deepfakes, *ACM Computing Surveys*, Jg. 54, H. 1, S. 1–41.

MIT Media Lab (Hg.) (n.n.), DeepFakes, Can You Spot Them?, 25.04.2023, https://detectfakes.media.mit.edu/.

Mittag, Gabriel/Sebastian Möller (2020), Deep Learning Based Assessment of Synthetic Speech Naturalness, in: *Interspeech 2020*, ISCA: ISCA, S. 1748–1752.

Mittal, Trisha/Uttaran Bhattacharya/Rohan Chandra/Aniket Bera/Dinesh Manocha (2020), Emotions Don't Lie, in: Chang Wen Chen (Hg.), *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY: ACM, S. 2823–2832.

Möhring, Wiebke/Daniela Schlütz (2019), *Die Befragung in der Medien- und Kommunikationswissenschaft. Eine praxisorientierte Einführung*, Wiesbaden: Springer Fachmedien Wiesbaden.

Montal, Tal/Zvi Reich (2017), I, Robot. You, Journalist. Who is the Author?, *Digital Journalism*, Jg. 5, H. 7, S. 829–849.

Muggli, Sandra (2014), *Im Netz ins Netz – Pädokriminalität im Internet und der Einsatz von verdeckten Ermittlern und verdeckten Fahndern zu deren Bekämpfung*, Genf/Zürich/Basel: SwissLex; Schulthess Juristische Medien AG.

Napolitano, Domenico (2020), The Cultural Origins of Voice Cloning, *Proceedings of the Eighth Conference on Computation, Communication, Aesthetics & X*.

Nelson, Jacob L./Harsh Taneja (2018), The small, disloyal fake news audience: The role of audience availability in fake news consumption, *New Media & Society*, Jg. 20, H. 10, S. 3720–3737.

Neto, Aguimar (2023), What is Latent Diffusion in AI?, 17.11.2023, https://medium.com/@aguimarneto/what-is-latent-diffusion-in-ai-43aa1ad4f71e.

Newman, Nic (2022), Digital News Report 2022, 16.05.2023, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf.

Newman, Nic/Richard Fletcher/Craig T. Robertson/Kirsten Eddy/Rasmus Kleis Nielsen (2022), Reuters Institute Digital News Report 2022, Oxford, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf.

Nguyen, Huy H./Junichi Yamagishi/Isao Echizen (2019), Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos, in: *ICASSP 2019*, Piscataway, NJ: IEEE, S. 2307–2311.

Nguyen, Thanh Thi/Quoc Viet Hung Nguyen/Dung Tien Nguyen/Duc Thanh Nguyen/Thien Huynh-The/Saeid Nahavandi/Thanh Tam Nguyen/Quoc-Viet Pham/Cuong M. Nguyen (2022), Deep learning for deepfakes creation and detection: A survey, *Computer Vision and Image Understanding*, Jg. 223, H. 103525, S. 1–19.

Nielsen, Rasmus Kleis (2017), The One Thing Journalism Just Might do for Democracy, *Journalism Studies*, Jg. 18, H. 10, S. 1251–1262.

Niessner, Gioia/Zanni/Bettina (2023), Glarner lässt Arslan mit KI gegen «kriminelle Türken» hetzen, 16.11.2023, https://www.fm1today.ch/schweiz/glarner-laesst-arslan-mit-ki-gegen-kriminelle-tuerken-hetzen-154236739.

Nightingale, Sophie J./Hany Farid (2022), AI-synthesized faces are indistinguishable from real faces and more trustworthy, *Proc. of NAS*, Jg. 119, H. 8.

Ning, Yishuang/Sheng He/Zhiyong Wu/Chunxiao Xing/Liang-Jie Zhang (2019), A Review of Deep Learning Based Speech Synthesis, *Applied Sciences*, Jg. 9, H. 19, S. 1–16.

Nobel, Peter/Rolf H. Weber (2021), *Medienrecht,* 4. Auflage, Bern: Stämpfli Verlag.

Nordenbrock, Kay (2022), Imagen Video: Auch Google präsentiert eine Video-KI, 17.11.2023, https://t3n.de/news/imagen-video-google-zeigt-video-ki-1503785/.

Nullmeier, Frank (2010), Strategie und politische Verwaltung. Anmerkungen zum Strategiepotential der Ministerialverwaltung, in: Joachim Raschke/Ralf Tils (Hg.), *Strategie in der Politikwissenschaft*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 257–265.

NZZ (2023), Cyberangriff auf das Unternehmen NZZ, *Neue Zürcher Zeitung*.

OECD (2022), Building Trust and Reinforcing Democracy. Preparing the Ground for Government Action, *OECD Public Governance Reviews*.

Oertel, Britta/Diego Dametto/Jakob Kluge/Jan Todt (2022), Algorithmen in digitalen Medien und ihr Einfluss auf die Meinungsbildung. Endbericht zum TA-Projekt.

Öhman, Carl/Luciano Floridi (2017), The Political Economy of Death in the Age of Information: A Critical Approach to the Digital Afterlife Industry, *Minds and Machines*, Jg. 27, H. 4, S. 639–662.

OpenAI (2022), DALL·E 2, 01.08.2022, https://openai.com/dall-e-2/.

OpenAI (2023), New AI classifier for indicating AI-written text, 18.08.2023, https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

OSCE (Hg.) (2017), Joint declaration on freedom of expression and “fake news”, disinformation and propaganda, 09.11.2023, https://www.osce.org/fom/302796.

Pallaske, Olaf (2022), Digitales Astroturfing: Wie unser Diskurs manipuliert wird, 25.04.2023, https://netzpolitik.org/2022/digitales-astroturfing-wie-unser-diskurs-manipuliert-wird/.

Paris, Britt/Joan Donovan (2019), Deepfakes and Cheap Fakes, *Data & Society*, S. 50.

Pawelec, Maria/Cora Bieß (2021), *Deepfakes*: Nomos Verlagsgesellschaft mbH & Co. KG.

PEN America (2022), Hard News: Journalists and the Threat of Disinformation, https://pen.org/report/hard-news-journalists-and-the-threat-of-disinformation/.

Peng, Xu/Bao Xintong (2022), An effective strategy for multi-modal fake news detection, *Multimedia Tools and Applications*, Jg. 81, H. 10, S. 13799–13822.

Peyravian, Mohammad/Allen Roginsky/Ajay Kshemkalyani (1998), On probabilities of hash value matches, *Computers & Security*, Jg. 17, H. 2, S. 171–176.

Pfefferkorn, Riana (2020), “Deepfakes” in the Courtroom, *Boston University Public Interest Law Journal*, Jg. 29, H. 2.

Pictory.ai (2023), Easy Video Creation For Content Marketers, 17.11.2023, https://pictory.ai/.

Porzelt, Benedikt (2015), Humor – Wie «heute show» und Co. unseren Blick auf Politik verändern (Benedikt Porzelt).

Posetti, Julie (2018a), Combatting online abuse: when journalists and their sources are targeted, in: Cherilyn Ireton/Julie Posetti (Hg.), *Journalism, ‘fake news’ & disinformation. Handbook for journalism education and training*, Paris, France: UNESCO, S. 115–127.

Posetti, Julie (2018b), News industry transformation: digital technology, social platforms and the spread of misinformation and disinformation, in: Cherilyn Ireton/Julie Posetti (Hg.), *Journalism, ‘fake news’ & disinformation. Handbook for journalism education and training*, Paris, France: UNESCO, S. 57–72.

Posetti, Julie/Nabeelah Shabbir/Diana Maynard/Kalina Bontcheva/Nermine Aboulez (2021), Global trends in online violence against women journalists, https://unesdoc.unesco.org/ark:/48223/pf0000377223.

Potter, W. James (2013), Review of Literature on Media Literacy: Media literacy, *Sociology Compass*, Jg. 7, H. 6, S. 417–435.

Prajwal, K. R./Rudrabha Mukhopadhyay/Vinay Namboodiri/C. V. Jawahar (2020), A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild.

Pu, Jiameng/Neal Mangaokar/Lauren Kelly/Parantapa Bhattacharya/Kavya Sundaram/Mobin Javed/Bolun Wang/Bimal Viswanath (2021), Deepfake Videos in the Wild: Analysis and Detection, in: Jure Leskovec (Hg.), *Proceedings of the WWW '21*: ACM, S. 981–992.

Puppis, Manuel (2019), Analyzing Talk and Text I: Qualitative Content Analysis, in: Hilde van den Bulck/Manuel Puppis/Karen Donders u.a. (Hg.), *The Palgrave Handbook of Methods for Media Policy Research*, Cham: Springer International Publishing, S. 367–384.

Puppis, Manuel/Michael Schenk/Brigitte Hofstetter (2017), *Medien und Meinungsmacht,* [1. Auflage], Zürich: vdf.

Raemy, Patric/Lea Hellmueller/Daniel Beck (2021), Journalists' contributions to political life in Switzerland: Professional role conceptions and perceptions of role enactment, *Journalism*, Jg. 22, H. 3, S. 767–786.

Raemy, Patric/Tim P. Vos (2021), A Negotiative Theory of Journalistic Roles, *Communication Theory*, Jg. 31, H. 1, S. 107–126.

Rahim, Zamira (2020), 'Deepfake' Queen delivers alternative Christmas speech, in warning about misinformation, 09.10.2022, https://edition.cnn.com/2020/12/25/uk/deepfake-queen-speech-christmas-intl-gbr/index.html.

Ramel, Raffael/André Vogelsang (2019), Kommentierung zu Art. 179novies ZGB, in: Marcel Alexander Niggli/Hans Wiprächtiger/Jürg-Beat Ackermann (Hg.), *Strafrecht. Strafgesetzbuch, Jugendstrafgesetz*, 4. Auflage, Basel: Helbing Lichtenhahn Verlag.

Rana, Md Shohel/Mohammad Nur Nobi/Beddhu Murali/Andrew H. Sung (2022), Deepfake Detection: A Systematic Literature Review, *IEEE Access*, Jg. 10, S. 25494–25513.

Rana, Md. Shohel/Andrew H. Sung (2020), DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection, in: *7th IEEE CSCloud and 6th IEEE EdgeCom*: IEEE, S. 70–75.

Razek, Ahmed (2018), Seeing isn't always believing, in: BBC, https://www.bbc.co.uk/blogs/aboutthebbc/entries/bee04c43-d896-4e36-8f02-244cb0db1c08.

Reber, Yannick (2020), Der neue Tatbestand des Identitätsmissbrauchs nach Art. 179decies StGB, *ex/ante*, Jg. 2020, H. 2, S. 33–37.

Rechsteiner, David/Christoph Errass (2014), Kommentierung zu Art. 16 BV, in: Bernhard Ehrenzeller/Banjamin Schindler/Rainer J. Schweizer u.a. (Hg.), *Die schweizerische Bundesverfassung. St. Galler Kommentar*, 3. Auflage, Genf/Zürich/Basel/St. Gallen: SwissLex; Dike Verlag AG.

Reddit (Hg.) (2020), Do not impersonate an individual or entity, 09.11.2023, https://support.reddithelp.com/hc/en-us/articles/360043075032.

Reddit (Hg.) (2023), Content Policy, 09.11.2023, https://www.redditinc.com/de-de/policies/content-policy.

Reh, Werner (1995), Quellen- und Dokumentenanalyse in der Politikfeldforschung: Wer steuert die Verkehrspolitik?, in: Ulrich Alemannkriz/Wolfgang Tonnesmannn/Volker Sommer (Hg.), *Politikwissenschaftliche Methoden. Grundriss für Studium und Forschung*, Olpaden: Westdeutscher Verlag, S. 201–259.

Riklin, Franz (2019a), Kommentierung zu Art. 173 StGB, in: Marcel Alexander Niggli/Hans Wiprächtiger/Jürg-Beat Ackermann (Hg.), *Strafrecht. Strafgesetzbuch, Jugendstrafgesetz*, 4. Auflage, Basel: Helbing Lichtenhahn Verlag.

Riklin, Franz (2019b), Kommentierung zu Art. 176 StGB, in: Marcel Alexander Niggli/Hans Wiprächtiger/Jürg-Beat Ackermann (Hg.), *Strafrecht. Strafgesetzbuch, Jugendstrafgesetz*, 4. Auflage, Basel: Helbing Lichtenhahn Verlag.

Rini, Regina/Leah Cohen (2022), Deepfakes, Deep Harms, *Journal of Ethics and Social Philosophy*, Jg. 22, H. 2.

Ritzi, Claudia (2017), Alte Sphären – neue Dimensionen. Die Politisierung von Privatheit im digitalen Zeitalter, S. 83–112.

Ritzi, Claudia/Alexandra Zierold (2019), Souveränität unter den Bedingungen der Digitalisierung, in: Isabelle Borucki/Wolf Jürgen Schünemann (Hg.), *Internet und Staat*: Nomos Verlagsgesellschaft mbH & Co. KG, S. 33–56.

Rivera, Gabriel (11.07.2023), AI may be‘running out of text in the universe’ to train chatbots, *Insider*, https://www.businessinsider.com/ai-could-run-out-text-train-chatbots-chatgpt-llm-2023-7, 18.08.2023.

Robin, Beglinger/Riedo Christof (2022), Ehrverletzungen im Internet – insbesondere auf Facebook, *medialex*.

Rogge, Marko (2018), Professionelle Angriffe auf Smartphones unter Zuhilfenahme von Social Engineering, *Wirtschaftsinformatik & Management*, Jg. 10, H. 5, S. 16–17.

Rohera, Dhiren/Harshal Shethna/Keyur Patel/Urvish Thakker/Sudeep Tanwar/Rajesh Gupta/Wei-Chiang Hong/Ravi Sharma (2022), A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects, *IEEE Access*, Jg. 10, S. 30367–30394.

Rohn, Patrick (2004), *Zivilrechtliche Verantwortlichkeit der Internet Provider nach schweizerischem Recht*, Genf/Zürich/Basel: SwissLex; Schulthess Juristische Medien AG.

Rosenthal, David (2014), Wie sich Privatpersonen gegen Verletzungen ihrer Persönlichkeits-Rechte durch Dritte auf Social-Media-Plattformen wehren können, *Anwaltsrevue*, H. 10, S. 415–422.

Rosner, Helen (7/17/2021), The Ethics of a Deepfake Anthony Bourdain Voice in "Roadrunner", *The New Yorker*, https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice, 16.11.2023.

Rossetti, Michael/Tauhid Zaman (2023), Bots, disinformation, and the first impeachment of U.S. President Donald Trump, *PLOS ONE*, Jg. 18, H. 5, e0283971.

RTS Info (2023), Les intelligences artificielles s'emparent de l'antenne de Couleur 3, in: Radio Télévision Suisse (RTS), 27.06.2023, https://www.rts.ch/info/sciences-tech/medias/13976194-les-intelligences-artificielles-semparent-de-lantenne-de-couleur-3.html.

Runway ML (2023), Gen-2, 17.11.2023, https://research.runwayml.com/gen2.

Sabir, Ekraam/Jiaxin Cheng/Ayush Jaiswal/Wael AbdAlmageed/Iacopo Masi/Prem Natarajan (2019), Recurrent Convolutional Strategies for Face Manipulation Detection in Videos, https://arxiv.org/pdf/1905.00582.

Salimans, Tim/Ian Goodfellow/Wojciech Zaremba/Vicki Cheung/Alec Radford/Xi Chen (2016), Improved Techniques for Training GANs, https://arxiv.org/pdf/1606.03498.

Saxena, Divya/Jiannong Cao (2022), Generative Adversarial Networks (GANs), *ACM Computing Surveys*, Jg. 54, H. 3, S. 1–42.

Schallbruch, Martin (2018), *Schwacher Staat im Netz*. *Wie die Digitalisierung den Staat in Frage stellt*, Wiesbaden: Springer.

Schapals, Aljosha Karim/Axel Bruns (2022), Responding to "Fake News": Journalistic Perceptions of and Reactions to a Delegitimising Force, *Media and Communication*, Jg. 10, H. 3.

Schenten, Ann-Kristin (27.06.2022), War der falsche Klitschko ein Deepfake oder Video-Schnittkunst?, *rbb24*, 25.04.2023.

Scherhag, Ulrich/Luca Debiasi/Christian Rathgeb/Christoph Busch/Andreas Uhl (2019), Detection of Face Morphing Attacks Based on PRNU Analysis, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Jg. 1, H. 4, S. 302–317.

Schifferes, Steve/Nic Newman/Neil Thurman/David Corney/Ayse Göker/Carlos Martin (2014), Identifying and Verifying News through Social Media, *Digital Journalism*, Jg. 2, H. 3, S. 406–418.

Schreiner, Maximilian (2022), Deepfakes: Wie alles begann - und wohin es führen könnte, 06.05.2022, https://mixed.de/geschichte-der-deepfakes-so-rasant-geht-es-mit-ki-fakes-voran/.

Schurter, Daniel (2022), «Cheap Fake» statt Deep-Fake – was wirklich hinter dem falschen Vitali Klitschko steckt, in: Watson, https://www.watson.ch/digital/ukraine/849274953-kein-deepfake-das-steckt-wirklich-hinter-dem-falschen-vitali-klitschko.

Sciforce (2/13/2020), Text-to-Speech Synthesis: an Overview - Sciforce - Medium, *Sciforce*, https://medium.com/sciforce/text-to-speech-synthesis-an-overview-641c18fcd35f, 06.05.2022.

Šepec, Miha (2020), Revenge Pornography or Non-Consensual Dissemination of Sexually Explicit Material as a Sexual Offence or as a Privacy Violation Offence, *International Journal of Cyber Criminology*, S. 418–438.

Sganga, Nicole (02.03.2022), Russia plotted to use fake video as pretense for Ukraine invasion, U.S. reveals, *CBS News*, https://www.cbsnews.com/news/russia-disinformation-video-ukraine-invasion-united-states/, 16.11.2023.

Shahzad, Sahibzada Adil/Ammarah Hashmi/Sarwar Khan/Yan-Tsung Peng/Yu Tsao/Hsin-Min Wang (2022), Lip Sync Matters: A Novel Multimodal Forgery Detector, in: *Proc. of APSIPA ASC*: IEEE, S. 1885–1892.

Sharir, Or/Barak Peleg/Yoav Shoham (2020), The Cost of Training NLP Models: A Concise Overview, http://arxiv.org/pdf/2004.08900v1.

Shawn, Shan/Wenger Emily/Zhang Jiayun/Li Huiying/Zheng Haitao/Y. Zhao Ben (2020), Fawkes: protecting privacy against unauthorized deep learning models, in: *Proceedings of the 29th USENIX Conference on Security Symposium*: USENIX Association, Article 90.

Shin, Soo Yun/Jiyoung Lee (2022), The Effect of Deepfake Video on News Credibility and Corrective Influence of Cost-Based Knowledge about Deepfakes, *Digital Journalism*, Jg. 10, H. 3, S. 412–432.

Shoemaker, Pamela J./Stephen D. Reese (2014), *Mediating the Message in the 21st Century*. *A Media Sociology Perspective*, Hoboken: Taylor and Francis.

Sidelov, Pavlo (2022), analysis of deepfakes problem for banks and financial institutions, *Veda a perspektivy*, H. 3.

Siegrist, Michael/Vivianne H. M. Visschers (2013), Acceptance of nuclear power: The Fukushima effect, *Energy Policy*, Jg. 59, S. 112–119.

Singer, Uriel/Adam Polyak/Thomas Hayes/Xi Yin/An Jie/Songyang Zhang/Qiyuan Hu/Harry Yang/Oron Ashual/Oran Gafni/Devi Parikh/Sonal Gupta/Yaniv Taigman (2022), Make-A-Video: Text-to-Video Generation without Text-Video Data, http://arxiv.org/pdf/2209.14792v1.

Slovic, Paul (1999), Trust, Emotion, Sex, Politics, and Science: Surveying the Risk-Assessment Battlefield, *Risk Analysis*, Jg. 19, H. 4, S. 689–701.

Smith, Nicola (13.07.2020), Taiwan builds 'nerd immunity' to resist Chinese disinformation campaigns, *The Telegraph*, https://www.telegraph.co.uk/news/2020/06/13/taiwan-builds-nerd-immunity-resist-chinese-disinformation-campaigns/, 16.11.2023.

Sohrawardi, Saniat J./Seng Sovantharith/Akash Chintha/Bao Thai/Raymond Ptutcha/Matthew Wright (2020), DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection, *USENIX Symposium on Usable Privacy and Security (SOUPS) 2020. August 9–11, 2020, Boston, MA, USA*.

Steinbach, Armin (2017), Meinungsfreiheit im postfaktischen Umfeld, *Juristen-Zeitung*, Jg. 72, H. 13, S. 653.

Stieler, Wolfgang (2021), Historische Fotos und Videos verbessern: Neuer Glanz für ganz alte Bilder, in: Heise Medien, 20.05.2022, https://www.heise.de/hintergrund/Neuer-Glanz-fuer-ganz-alte-Bilder-5027502.html.

Störk-Biber, Constanze/Jürgen Hampel/Cordula Kropp/Michael Zwick (2020), Wahrnehmung von Technik und Digitalisierung in Deutschland und Europa: Befunde aus dem TechnikRadar, *HMD Praxis der Wirtschaftsinformatik*, Jg. 57, H. 1, S. 21–32.

Strasser, Matthias (2023), Regulierung von KI - Mitte-links-Allianz: Keine Deepfakes im Wahlkampf, 16.11.2023, https://www.srf.ch/news/schweiz/wahlen-2023/regulierung-von-ki-mitte-links-allianz-keine-deepfakes-im-wahlkampf.

Strub, Jean-Daniel/Francesca Bosisio/Ralf J. Jox/Anca-Cristina Sterie (2024), *La mort à l'ère numérique*. *Chances et risques du Digital Afterlife,* 1. Auflage, Zollikon: vdf Hochschulverlag AG.

Sun, Pu/Yuezun Li/Honggang Qi/Siwei Lyu (2022a), Faketracer: Exposing Deepfakes with Training Data Contamination, in: *2022 IEEE International Conference on Image Processing (ICIP)*: IEEE, S. 1161–1165.

Sun, Yasheng/Hang Zhou/Kaisiyuan Wang/Qianyi Wu/Zhibin Hong/Jingtuo Liu/Errui Ding/Jingdong Wang/Ziwei Liu/Koike Hideki (2022b), Masked Lip-Sync Prediction by Audio-Visual Contextual Exploitation in Transformers, in: Soon Ki Jung/Jehee Lee/Adam Bargteil (Hg.), *Proc. of SA '22: SIGGRAPH Asia*, NY, USA: ACM, S. 1–9.

Suvorova, Inna (2022), Deepfake pornography as a male gaze on fan culture, https://arxiv.org/pdf/2202.00374.

Swico (Hg.) (2013), SIMSA: Code of Conduct, 11.12.2023, https://haerting.ch/wissen/simsa-code-of-conduct-verhaltensanweisungen-fuer-hosting-provider/.

Sylvester, Shannon (2021), Don't Let Them Fake You Out: How Artificially Mastered Videos Are Becoming the Newest Threat in the Disinformation War and What Social Media Platforms Should Do About It, *Federal Communications Law Journal*, H. 73, S. 369–392.

Symantec Threat Hunter Team (2021), The Threat Landscape in 2021 (White Paper). White Paper.

SynSIG (Hg.) (2022), Blizzard Challenge, 02.08.2022, https://www.synsig.org/index.php/Blizzard_Challenge.

Tag, Brigitte/Martin Wyss, Die strafrechtliche Einordnung von pornografischen Deepfakes, in: *Jusletter*, 29. April 2024, Rz. 30.

Tal-Or, Nurit/Jonathan Cohen/Yariv Tsfati/Albert C. Gunther (2010), Testing Causal Direction in the Influence of Presumed Media Influence, *Communication Research*, Jg. 37, H. 6, S. 801–824.

Tandoc, Edson C./Joy Jenkins/Stephanie Craft (2019), Fake News as a Critical Incident in Journalism, *Journalism Practice*, Jg. 13, H. 6, S. 673–689.

Tandoc, Edson C., Jr./Andrew Z. H. Yee/Jeremy Ong/James Chong Boi Lee/Duan Xu/Zheng Han/Chew Chee Han Matthew/Ng, Janelle Shaina Hui Yi/Cui Min Lim/Lydia Rui Jun Cheng/Marie Ysa Cayabyab (2021), Developing a Perceived Social Media Literacy Scale: Evidence from Singapore, *International Journal of Communication*, Jg. 15, S. 2484–2505.

Tangermann, Victor (2023), There's a Problem With That App That Detects GPT-Written Text: It's Not Very Accurate, 18.08.2023, https://futurism.com/gptzero-accuracy.

Tay, Siok Wah/Pin Shen Teh/Stephen J. Payne (2021), Reasoning about privacy in mobile application install decisions: Risk perception and framing, *International Journal of Human-Computer Studies*, Jg. 145, S. 102517.

Taylor, Luke (05.02.2023), Amnesty International criticised for using AI-generated images, *The Guardian*, https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism, 16.05.2023.

TechRepublic (2021), Deepfakes: Microsoft and others in big tech are working to bring authenticity to videos, photos, 13.05.2022, https://www.techrepublic.com/article/deepfakes-microsoft-and-others-in-big-tech-are-working-to-bring-authenticity-to-videos-photos/.

Temir, Erkam (2020), Deepfake: New Era in The Age of Disinformation & End of Reliable Journalism, *SELÇUK İLETİŞİM DERGİSİ*, Jg. 13, H. 2.

Ternovski, John/Joshua Kalla/Peter Michael Aronow (2021), Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments.

The New York State Senate (Hg.) (2020), 2019-S5959D, 09.11.2023, https://www.nysenate.gov/legislation/bills/2019/S5959.

Thies, Justus/Michael Zollhofer/Marc Stamminger/Christian Theobalt/Matthias Nießner (2016), Face2face: Real-time face capture and reenactment of rgb videos, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, S. 2387–2395.

TikTok (Hg.) (2023), Integrität und Authentizität, 09.11.2023, https://www.tiktok.com/community-guidelines/de-de/integrity-authenticity/.

Toler, Aric (2019), Amazon's Online Bezos Brigade Unleashed On Twitter, 17.05.2023, https://www.bellingcat.com/news/americas/2019/08/15/amazons-online-bezos-brigade-unleashed-on-twitter/.

Tribelhorn, Marc (7/30/2022), Die Mär von der Polarisierung – oder: die Schweiz, das Dromedar, *Neue Zürcher Zeitung*, https://www.nzz.ch/meinung/die-maer-von-der-polarisierung-oder-die-schweiz-das-dromedar-ld.1695768, 16.05.2023.

Tschannen, Pierre (2021), *Staatsrecht der Schweizerischen Eidgenossenschaft,* 5. Auflage, Bern: Stämpfli Verlag.

Tsfati, Yariv/H. G. Boomgaarden/J. Strömbäck/R. Vliegenthart/A. Damstra/E. Lindgren (2020), Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis, *Annals of the International Communication Association*, Jg. 44, H. 2, S. 157–173.

Turnitin (2023), Empower Students to Do Their Best, Original Work, 18.08.2023, https://www.turnitin.com/.

University of Baltimore (Hg.) (2020), Prof. Cavazos's Latest Report: Digital Ad Fraud Costs Will Rise to $35B Globally This Year, 16.05.2023, https://www.ubalt.edu/news/news-releases.cfm?id=3621.

UVEK/EDA/EFD/WBF (2019), Die Schweiz und der digitale Binnenmarkt der EU, 06.11.2023, https://www.bakom.admin.ch/dam/bakom/de/dokumente/informationsgesellschaft/mise%20en%20oeuvre/DSM_EU-CH.pdf.download.pdf/Die%20Schweiz%20und%20der%20digitale%20Binnenmarkt%20der%20EU%20-%2014-06-2019.pdf.

Vaccari, Cristian/Andrew Chadwick (2020), Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News, *Social Media + Society*, Jg. 6, H. 1, 205630512090340.

van Aelst, Peter/Jesper Strömbäck/Toril Aalberg/Frank Esser/Claes de Vreese/Jörg Matthes/David Hopmann/Susana Salgado/Nicolas Hubé/Agnieszka Stępińska/Stylianos Papathanassopoulos/Rosa Berganza/Guido Legnante/Carsten Reinemann/Tamir Sheafer/James Stanyer (2017), Political communication in a high-choice media environment: a challenge for democracy?, *Annals of the International Communication Association*, Jg. 41, H. 1, S. 3–27.

van der Sloot, Bart/Yvette Wagensveld (2021), Deepfakes: The legal challenges of a synthetic society, https://www.tilburguniversity.edu/sites/default/files/download/Deepfake%20EN.pdf.

van der Sloot, Bart/Yvette Wagensveld (2022), Deepfakes: regulatory challenges for the synthetic society, *Computer Law & Security Review*, Jg. 46.

van der Wilk, Adriane (2021), *Protecting women and girls from violence in the digital age*. *The relevance of the Istanbul Convention and the Budapest Convention on Cybercrime in addressing online and technology-facilitated violence against women*, Strasbourg: Council of Europe.

van Huijstee, Mariëtte/Pieter van Boheemen/Djurre Das/Linda Nierling/Jutta Jahnel/Murat Karaboga/Martin Fatun (2021), Tackling deepfakes in European policy, 06.05.2022, https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)690039.

Veerasamy, Namosha/Heloise Pieterse (2022), Rising Above Misinformation and Deepfakes, *International Conference on Cyber Warfare and Security*, Jg. 17, H. 1, S. 340–348.

Verhulst, Stefaan (2020), Malicious Uses and Abuses of Artificial Intelligence, 16.05.2023, https://thelivinglib.org/malicious-uses-and-abuses-of-artificial-intelligence/.

Verwaltung innovativ (2023), 3.10 Risikomanagement, in: BMI, 26.04.2023, https://www.verwaltung-innovativ.de/OHB/DE/Organisationshandbuch-NEU/3_managementansaetze_u_instrumente/3_10_Risikomanagement/risikomanagement-node.html.

vicomtech (2023), Voice Cloning. Markets, 22.09.2023, https://www.speechandlanguagesolutions.com/en/voice-cloning/.

Vizoso, Ángel/Martín Vaz-Álvarez/Xosé López-García (2021), Fighting Deepfakes: Media and Internet Giants' Converging and Diverging Strategies Against Hi-Tech Misinformation, *Media and Communication*, Jg. 9, H. 1, S. 291–300.

Vu, Hong Tien/Magdalena Saldaña (2021), Chillin' Effects of Fake News: Changes in Practices Related to Accountability and Transparency in American Newsrooms Under the Influence of Misinformation and Accusations Against the News Media, *Journalism & Mass Communication Quarterly*, 107769902098478.

Wagner, Hans/Philomen Schönhagen (2021), *Qualitative Methoden der Kommunikationswissenschaft,* 3., aktualisierte Auflage, Baden-Baden: Nomos.

Wagner, Travis L./Ashley Blewer (2019), "The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video, *Open Information Science*, Jg. 3, H. 1, S. 32–46.

Wahl-Jorgensen, Karin/Matt Carlson (2021), Conjecturing Fearful Futures: Journalistic Discourses on Deepfakes, *Journalism Practice*, S. 1–18.

Waisbord, Silvio (2018), Truth is What Happens to News, *Journalism Studies*, Jg. 19, H. 13, S. 1866–1878.

Walsh, Joseph (12/16/2016), Rogue One: the CGI resurrection of Peter Cushing is thrilling – but is it right?, *The Guardian*, https://www.theguardian.com/film/filmblog/2016/dec/16/rogue-one-star-wars-cgi-resurrection-peter-cushing, 16.11.2023.

Wang, Run/Felix Juefei-Xu/Lei Ma/Xiaofei Xie/Yihao Huang/Jian Wang/Yang Liu (2019), FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces, https://arxiv.org/pdf/1909.06122.

Wang, Soyoung/Seongcheol Kim (2022), Users' emotional and behavioral responses to deepfake videos of K-pop idols, *Computers in Human Behavior*, Jg. 134, S. 107305.

Wang, Zuoguang/Hongsong Zhu/Limin Sun (2021), Social Engineering in Cybersecurity: Effect Mechanisms, Human Vulnerabilities and Attack Methods, *IEEE Access*, Jg. 9, S. 11895–11910.

Wardle, Claire/Hossein Derakhshan (2018), Information disorder: Toward an interdisciplinary framework for research and policy making, 17.11.2023, https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

Wenger, Emily/Max Bronckers/Christian Cianfarani/Jenna Cryan/Angela Sha/Haitao Zheng/Ben Y. Zhao (2021), "Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World, in: Yongdae Kim (Hg.), *Proceedings of the 2021 ACM SIGSAC*: ACM, S. 235–251.

Westerlund, Mika (2019), The Emergence of Deepfake Technology: A Review, *Technology Innovation Management Review*, Jg. 9, H. 11, S. 39–52.

Westling, Jeffrey (2019), Are Deep Fakes a Shallow Concern?, TPRC47: The 47th Research Conference on Communication, Information and Internet Policy 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3426174.

Whittaker et al. (2020), Around Me Are Synthetic Faces: The Mad World of AI-Generated Media.

Widmer, Michèle (2023), M Le Média: Avatar moderiert auf Westschweizer TV-Sender, in: Persönlich.com, 27.06.2023, https://www.persoenlich.com/medien/avatar-moderiert-auf-westschweizer-tv-sender.

Wiegand, Dorothee (2023), *Synthetische Stimme verschafft Zugang zu Bankdaten*, *heise online*, https://www.heise.de/news/Synthetische-Stimme-verschafft-Zugang-zu-Bankdaten-7527666.html, 19.03.2023.

Williams, Daniel (7.7.2023), The Fake News about Fake News. In Foolproof, psychologist Sander van der Linden compares misinformation to viral infection—and claims to have a vaccine, *Boston Review*, https://www.bostonreview.net/articles/the-fake-news-about-fake-news/.

Wilson, Rose J./Pauline Paterson/Caitlin Jarrett/Heidi J. Larson (2015), Understanding factors influencing vaccination acceptance during pregnancy globally: A literature review, *Vaccine*, Jg. 33, H. 47, S. 6420–6429.

Wöbbeking, Jan Philipp (2023), KI-Videogenerator Kaiber: Animationen aus Text und Videos erzeugen, in: Heise Medien, 17.11.2023, https://www.heise.de/ratgeber/KI-Videogenerator-Kaiber-Animationen-aus-Text-und-Videos-erzeugen-9349841.html.

Wojewidka, John (2020), The deepfake threat to face biometrics, *Biometric Technology Today*, Jg. 2020, H. 2, S. 5–7.

Wyttenbach, Judith (2015), Kommentierung zu Art. 21 BV, in: Bernhard Waldmann/Eva Maria Belser/Astrid Epiney (Hg.), *Bundesverfassung*, Basel: Helbing-Lichtenhahn-Verl.

Yadlin-Segal, Aya/Yael Oppenheim (2021), Whose dystopia is it anyway? Deepfakes and social media regulation, *Convergence: The International Journal of Research into New Media Technologies*, Jg. 27, H. 1, S. 36–51.

Yar, Majid (2005), The Novelty of 'Cybercrime': An Assessment in Light of Routine Activity Theory, *European Journal of Criminology*, Jg. 2, H. 4, S. 407–427.

Yazdinejad, Abbas/Reza M. Parizi/Gautam Srivastava/Ali Dehghantanha (2020), Making Sense of Blockchain for AI Deepfakes Technology, in: *Proc. of IEEE GC Wkshps*, Piscataway, NJ: IEEE, S. 1–6.

Yu, Ning/Larry Davis/Mario Fritz (2018), Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints, https://arxiv.org/pdf/1811.08180.

Zeller, Franz (2004), *Öffentliches Medienrecht*. *Mit einer Kurzeinführung in die Rechtswissenschaft*, Bern: Stämpfli.

Zhou, Yipin/Ser-Nam Lim (2021), Joint Audio-Visual Deepfake Detection, in: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*: IEEE, S. 14780–14789.

Zimmermann, Fabian/Matthias Kohring (2018), «Fake News» als aktuelle Desinformation. Systematische Bestimmung eines heterogenen Begriffs, *Medien & Kommunikationswissenschaft*, Jg. 66, H. 4, S. 526–541.

Zoss, Gaspard/Prashanth Chandran/Eftychios Sifakis/Markus Gross/Paulo Gotardo/Derek Bradley (2022), Production-Ready Face Re-Aging for Visual Effects, *ACM Transactions on Graphics*, Jg. 41, H. 6, S. 1–12.

Zwahlen, Fabienne/Irene Marti/Marina Richter/Cathrine Konopatsch/Ueli Hostettler (2020), Wirtschaftsspionage in der Schweiz – Schlussbericht zuhanden des Nachrichtendienstes des Bundes (NDB), Bern.

# Anhang

## A.1. Detaillierte Abbildungen der Wahrnehmungsstudie

**Tabelle 17:** Lineares Regressionsmodell für Wahrnehmung von Risiken für die Politik

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| (Intercept) | 3.06 | 2.39 – 3.74 | < 0.001 |
| Einfluss auf eigene Meinung | − 0.04 | − 0.20 – 0.12 | 0.64 |
| Einfluss auf Meinung anderer | 0.42 | 0.32 – 0.52 | < 0.001 |
| Verbreitung von Deepfakes | 0.12 | 0.05 – 0.20 | 0.002 |
| Erfahrung mit Deepfakes | 0.08 | − 0.07 – 0.22 | 0.285 |
| Deepfake Exposure | − 0.01 | − 0.07 – 0.05 | 0.766 |
| Deepfakes erkennen | − 0.03 | − 0.08 – 0.02 | 0.273 |
| Vertrauen in Institutionen | − 0.08 | − 0.15 – − 0.01 | 0.023 |
| Alter | 0 | − 0.00 – 0.00 | 0.838 |
| Sprachregion [FR] | − 0.46 | − 0.62 – − 0.29 | < 0.001 |
| Bildung [Hochschulabschluss] | − 0.02 | − 0.20 – 0.16 | 0.82 |
| Gender | − 0.07 | − 0.24 – 0.09 | 0.371 |
| Einfluss eigene * andere Meinung | 0.01 | − 0.02 – 0.04 | 0.455 |
| Fälle | 1361 | | |
| $R^2$ / $R^2$ korrigiert | 0.229 / 0.222 | | |

**Tabelle 18:** Lineares Regressionsmodell für Wahrnehmung von Risiken für die Medien

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| (Intercept) | 3.62 | 3.11 – 4.12 | < 0.001 |
| Einfluss auf eigene Meinung | − 0.05 | − 0.17 – 0.07 | 0.452 |
| Einfluss auf Meinung anderer | 0.3 | 0.23 – 0.38 | < 0.001 |
| Verbreitung von Deepfakes | 0.18 | 0.12 – 0.23 | < 0.001 |
| Erfahrung mit Deepfakes | 0.08 | − 0.03 – 0.19 | 0.161 |
| Deepfake Exposure | − 0.03 | - 0.07 – 0.02 | 0.239 |
| Deepfakes erkennen | − 0.07 | − 0.10 – − 0.03 | 0.001 |
| Vertrauen in Institutionen | 0.06 | 0.01 – 0.12 | 0.014 |
| Alter | 0 | − 0.00 – 0.00 | 0.994 |
| Sprachregion [FR] | − 0.2 | − 0.32 – − 0.07 | 0.002 |
| Bildung [Hochschule] | 0.03 | − 0.11 – 0.16 | 0.713 |
| Gender | 0.07 | − 0.05 – 0.19 | 0.236 |
| Einfluss eigene * andere Meinung | 0 | − 0.02 – 0.02 | 0.881 |
| Fälle | 1361 | | |
| $R^2$ / $R^2$ korrigiert | 0.198 / 0.191 | | |

**Tabelle 19:** Lineares Regressionsmodell für Wahrnehmung von Risiken für die Wirtschaft

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| (Intercept) | 3.17 | 2.56 – 3.78 | < 0.001 |
| Einfluss auf eigene Meinung | 0.01 | − 0.13 – 0.16 | 0.858 |
| Einfluss auf Meinung anderer | 0.39 | 0.30 – 0.48 | < 0.001 |
| Verbreitung von Deepfakes | 0.04 | − 0.03 – 0.11 | 0.229 |
| Erfahrung mit Deepfakes | − 0.07 | − 0.20 – 0.06 | 0.281 |
| Deepfake Exposure | − 0.03 | − 0.08 – 0.02 | 0.282 |
| Deepfakes erkennen | − 0.02 | − 0.07 – 0.03 | 0.405 |

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| Vertrauen in Institutionen | − 0.01 | − 0.07 – 0.06 | 0.858 |
| Alter | 0 | − 0.00 – 0.00 | 0.123 |
| Sprachregion [FR] | − 0.11 | − 0.26 – 0.04 | 0.158 |
| Bildung [Hochschule] | − 0.18 | − 0.34 – − 0.01 | 0.033 |
| Gender | − 0.07 | − 0.21 – 0.08 | 0.386 |
| Einfluss eigene * andere Meinung | 0.01 | − 0.02 – 0.03 | 0.706 |
| Fälle | 1361 | | |
| $R^2$ / $R^2$ korrigiert | 0.198 / 0.191 | | |

**Tabelle 20:** Lineares Regressionsmodell für Wahrnehmung von individuellen Risiken

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| (Intercept) | 2.23 | 1.43 – 3.03 | < 0.001 |
| Einfluss auf eigene Meinung | 0.14 | − 0.06 – 0.33 | 0.168 |
| Einfluss auf Meinung anderer | 0.17 | 0.05 – 0.29 | 0.005 |
| Verbreitung von Deepfakes | 0.05 | − 0.04 – 0.14 | 0.304 |
| Erfahrung mit Deepfakes | − 0.06 | − 0.23 – 0.11 | 0.511 |
| Deepfake Exposure | 0.03 | − 0.03 – 0.10 | 0.329 |
| Deepfakes erkennen | 0.02 | − 0.04 – 0.08 | 0.47 |
| Vertrauen in Institutionen | − 0.1 | − 0.18 – − 0.02 | 0.015 |
| Alter | 0 | − 0.00 – 0.00 | 0.306 |
| Sprachregion [FR] | − 0.09 | − 0.29 – 0.11 | 0.379 |
| Bildung [Hochschule] | − 0.29 | − 0.50 – − 0.08 | 0.007 |
| Gender | 0.31 | 0.11 – 0.50 | 0.002 |
| Einfluss eigene * andere Meinung | 0.01 | − 0.02 – 0.05 | 0.465 |
| Fälle | 1361 | | |
| $R^2$ / $R^2$ korrigiert | 0.130 / 0.123 | | |

**Tabelle 21:** Lineares Regressionsmodell für Wahrnehmung von Chancen für die Medien

| **Prädiktoren** | **Estimates** | **CI** | **p** |
|---|---|---|---|
| (Intercept) | 3.54 | 2.99 – 4.09 | < 0.001 |
| Label [synthetische Medien] | 0.43 | 0.20 – 0.66 | < 0.001 |
| Alter | − 0.01 | − 0.02 – − 0.01 | 0.001 |
| Gender | 0.43 | 0.20 – 0.66 | < 0.001 |
| Sprachregion [FR] | − 0.07 | − 0.34 – 0.20 | 0.613 |
| Bildung [Hochschule] | 0.1 | − 0.15 – 0.36 | 0.43 |
| Fälle | 660 | | |
| $R^2$ / $R^2$ korrigiert | 0.058 / 0.051 | | |

**Tabelle 22:** Lineares Regressionsmodell für Wahrnehmung von Chancen für die Wirtschaft

| **Predictors** | **Estimates** | **CI** | **p** |
|---|---|---|---|
| (Intercept) | 4.01 | 3.54 – 4.49 | < 0.001 |
| Label [synthetische Medien] | 0.36 | 0.16 – 0.55 | < 0.001 |
| Alter | − 0.01 | − 0.02 – − 0.01 | < 0.001 |
| Gender | 0.13 | − 0.07 – 0.33 | 0.197 |
| Sprachregion [FR] | − 0.31 | − 0.54 – − 0.07 | 0.012 |
| Bildung [Hochschule] | 0.21 | − 0.01 – 0.44 | 0.063 |
| Fälle | 660 | | |
| $R^2$ / $R^2$ korrigiert | 0.057 / 0.050 | | |

**Tabelle 23:** Lineares Regressionsmodell für Wahrnehmung von individuellen Chancen

| Predictors | Estimates | CI | p |
|---|---|---|---|
| (Intercept) | 4.31 | 3.78 – 4.84 | < 0.001 |
| Label [synthetische Medien] | 0.21 | − 0.01 – 0.43 | 0.060 |
| Alter | − 0.02 | − 0.03 – − 0.01 | < 0.001 |
| Gender | − 0.02 | − 0.24 – 0.20 | 0.858 |
| Sprachregion [FR] | − 0.05 | − 0.31 – 0.22 | 0.730 |
| Bildung [Hochschule] | 0.22 | − 0.03 – 0.47 | 0.086 |
| Fälle | 660 | | |
| $R^2$ / $R^2$ korrigiert | 0.052 / 0.045 | | |

**Tabelle 24:** Mehrebenen-Regressionsmodell mit varying intercepts für Video (n = 6) und Teilnehmenden (n = 1361). Erkennungskompetenz für Teilnehmende mit und ohne Literacy-Intervention.

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| (Intercept) | 4.14 | 3.68 – 4.60 | < 0.001 |
| Videotyp [real] | − 0.56 | − 1.43 – 0.31 | 0.208 |
| Literacy Intervention | − 0.05 | − 0.15 – 0.06 | 0.373 |
| Alter | 0 | − 0.00 – 0.00 | 0.204 |
| Sprachregion [FR] | − 0.12 | − 0.24 – − 0.01 | 0.032 |
| Bildung [Hochschule] | 0.13 | 0.01 – 0.25 | 0.029 |
| Gender | − 0.05 | − 0.16 – 0.06 | 0.331 |
| Bekanntheit Person in Video | − 0.02 | − 0.04 – 0.01 | 0.159 |
| Videotyp [real]* Literacy Intervention | − 0.06 | − 0.23 – 0.10 | 0.451 |
| Videotyp [real]* Alter | 0 | − 0.00 – 0.00 | 0.387 |
| Videotyp [real]* Sprachregion [FR] | 0.06 | − 0.12 – 0.24 | 0.521 |

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| Videotyp [real]* Bildung [Hochschule] | − 0.2 | − 0.39 – − 0.01 | 0.037 |
| Videotyp * Gender | 0.11 | − 0.06 – 0.28 | 0.208 |
| Videotyp [real] * Bekanntheit | − 0.07 | − 0.11 – − 0.02 | 0.003 |

| Random Effects | |
|---|---|
| Mass | Wert |
| σ2 | 3.58 |
| τ00 Teilnehmer | 0.36 |
| τ00 Video | 0.24 |
| ICC | 0.14 |
| N Teilnehmer | 1361 |
| N Video | 6 |
| Fälle | 8166 |
| Marginales $R^2$ / Konditionales $R^2$ | 0.035 / 0.173 |

| Prädiktoren | Estimates | CI | p |
|---|---|---|---|
| Intercept | 3.65 | 2.94 – 4.37 | < 0.001 |
| Videotyp [real] | − 0.24 | − 1.47 – 0.98 | 0.697 |
| Social Media Literacy | 0.1 | 0.03 – 0.17 | 0.005 |
| Internet−Skills | 0.04 | − 0.02 – 0.10 | 0.158 |
| Medienkompetenz | − 0.05 | − 0.14 – 0.05 | 0.345 |
| Alter | 0 | − 0.00 – 0.00 | 0.356 |
| Region [FR] | − 0.05 | − 0.20 – 0.10 | 0.489 |
| Bildung [Hochschule] | 0.04 | − 0.12 – 0.20 | 0.596 |
| Gender | − 0.06 | − 0.21 – 0.10 | 0.454 |
| Erfahrung mit Deepfakes | 0.11 | − 0.02 – 0.25 | 0.096 |

| **Prädiktoren** | **Estimates** | **CI** | **p** |
|---|---|---|---|
| Exposure zu Deepfakes | − 0.03 | − 0.08 – 0.02 | 0.19 |
| Videotyp [real] * Social−Media−Literacy | − 0.22 | − 0.34 – − 0.11 | < 0.001 |
| Videotyp [real] * Internet−Skills | − 0.07 | − 0.16 – 0.03 | 0.172 |
| Videotyp [real] * Medienkompetenz | 0.15 | − 0.00 – 0.30 | 0.057 |
| Videotyp [real] * Alter | 0 | − 0.00 – 0.00 | 0.718 |
| Videotyp [real] * Region [FR] | − 0.11 | − 0.35 – 0.14 | 0.387 |
| Videotyp [real] * Bildung [Hochschule] | − 0.04 | − 0.30 – 0.21 | 0.741 |
| Videotyp [real] * Gender | 0.1 | − 0.15 – 0.35 | 0.418 |
| Videotyp [real] * Erfahrung | − 0.1 | − 0.31 – 0.12 | 0.386 |
| Videotyp [real] * Exposure | 0.05 | − 0.04 – 0.13 | 0.27 |

| Random Effects | |
|---|---|
| Mass | Wert |
| σ2 | 3.37 |
| τ00 Teilnehmer | 0.29 |
| τ00 Video | 0.18 |
| ICC | 0.12 |
| N Teilnehmer | 691 |
| N Video | 6 |
| Fälle | 4146 |
| Marginales $R^2$ / Konditionales $R^2$ | 0.046 / 0.163 |

## A.2. Interviewleitfaden TA-SWISS-Projekt

Thema: Deepfakes im Journalismus

*Einstieg*

Video als Beispiel für Deepfakes

Was sind Ihre Gedanken zu diesem Video in Bezug auf die journalistische Arbeit?

Wurden Sie in Ihrer Arbeit bereits einmal mit Deepfakes konfrontiert oder kennen Sie Kolleg:innen, bei denen das der Fall war? Wie sind Sie damit umgegangen?

*Potenziale und Risiken der Technologien hinter Deepfakes*

Wir würden nun gerne etwas spezifischer auf die Potenziale und Risiken von Deepfakes für den Journalismus und Medienorganisationen eingehen.

Welche Potenziale der Technologie sehen Sie für den Journalismus oder Medienorganisationen heute oder künftig?

Kommen wir nun zu den Risiken. Welche Risiken der Technologie sehen Sie für den Journalismus oder Medienorganisationen heute oder künftig?

*Prävention innerhalb der Redaktion/des Medienhauses*

Nun möchten wir gerne noch etwas über mögliche Präventionsmassnahmen im Medienunternehmen sprechen.

Wie werden Deepfakes redaktionsintern resp. medienorganisationsintern diskutiert?

Wie wird das Bewusstsein der Journalist:innen für Deepfakes intern gefördert?

Gibt es Kooperationen zwischen Medienunternehmen und Forschung oder spezifische Trainings zur Verringerung der Risiken von Deepfakes? (z.B. Trainingsdatensätze oder gemeinsame Datenbanken).

Werden die Mitarbeiter:innen ermutigt, externe Kurse zu diesem Thema zu besuchen?

*Erkennung und Filterung von Deepfakes*

Da die Technologien hinter Deepfakes komplex sind, sind diese auch schwieriger zu erkennen als sonstige Desinformation oder Fake News. Kommen wir also zu den Massnahmen zur Erkennung von Deepfakes in Ihrer Redaktion. Dabei

können wir gerne über bereits bestehende oder auch über derzeit diskutierte Massnahmen sprechen.

Welche Strategien gibt es, um die Veröffentlichung resp. Berücksichtigung von Deepfakes in der eigenen Berichterstattung zu vermeiden?

Gibt es z.B. interne Regeln zum Vorgehen der Prüfung auf Deepfakes.

Gibt es spezialisierte Abteilungen im Medienhaus?

Werden Technologien und Software zur Erkennung von Deepfakes eingesetzt (z.B. Algorithmen/Machine Learning, Wasserzeichen, Blockchain-basierte Scanner)?

*Massnahmen, wenn Journalist:innen betroffen sind*

Interviewer legt Artikel zu Fall in Indien vor, wo eine Journalistin mittels eines Deepfakes erpresst wurde.

Was sind Ihre Gedanken zu diesem Fall, wenn Sie an Ihre eigene Arbeit als Journalist:in denken?

Welche Strategien gibt es in Ihrer Medienorganisation, falls Journalist:innen selbst Opfer von Deepfakes werden? (Vorbereitung, Rechtsschutz und Notfallpläne)

*Intervention und Reparation bei der Veröffentlichung von Deepfakes*

Kommen wir nun zu den Interventions- und Reparationsmöglichkeiten, für den Fall, dass ein Medium tatsächlich unwillentlich Deepfakes verbreitet.

Welche Notfallstrategien gibt es bei einer Veröffentlichung von Deepfakes?

Wie wird dafür gesorgt, dass nach der Publikation eines Deepfakes eine Korrektur von den Nutzer:innen zur Kenntnis genommen wird?

*Sensibilisierung und Förderung von Medienkompetenz beim Publikum*

Einerseits können Medienorganisationen Massnahmen ergreifen, andererseits aber auch dafür sorgen, dass Nutzer:innen kompetent mit Inhalten umgehen können. Uns beschäftigt deshalb auch, wie eine Sensibilisierung und Förderung der Medienkompetenz beim Publikum gelingen kann.

Inwiefern sind Deepfakes auch ein Thema für die eigene Berichterstattung, um die Nutzer:innen dafür zu sensibilisieren?

Und jetzt allgemein/nicht nur Deepfakes: Welche Rolle sollten Medien in der Vermittlung von Medienkompetenz und speziell Informationskompetenz spielen?

Welche Chancen bietet die Vermittlung von Medienkompetenz oder speziell Informationskompetenz für den Journalismus insgesamt oder für Ihre Medienorganisation im Speziellen?

*Abschluss*

Gibt es Aspekte, die wir bisher nicht angesprochen haben, die Ihnen aber in Zusammenhang mit Deepfakes und Desinformation wichtig erscheinen?

## A.3. Liste der Codes

| |
|---|
| Vertrauen in die Medien |
| Shallow Fakes als grössere Bedrohung |
| Risiken von Deepfakes |
| Aktuell noch nicht die grosse Gefahr |
| Konfrontation mit Deepfakes im journ. Alltag |
| Deepfakes werden nicht erkannt und verbreitet |
| User-generierte Deepfakes als Kommentare |
| Stellen Akzeptanz und Glaubwürdigkeit infrage |
| Bürger stellen jede Wahrheit infrage |
| Schweiz als Schonraum für Deepfakes |
| Medieninhalte werden für Deepfakes missbraucht |
| Chancen von synthetischen Medien für Journalismus |
| Kritische Sicht auf Chancen |
| Grenze zwischen Deepfakes und synthetischen Medien |
| Verdeckte Recherche |
| Arbeitsgruppen und Entwicklungen im Medienhaus |
| Technologie ist noch nicht reif genug |
| Newsbeiträge attraktiver gestalten |
| Instrumente und Abteilungen aufbauen und vermarkten |
| Erfolgreicher Umgang mit Deepfakes als Qualitätsmerkmal |
| Personalisierung, Avatare und virtuelle Räume |
| Automatisierung in mehreren Sprachen |
| Tote Personen sprechend in Videos darstellen |
| Sensibilisierung der Journalist:innen |
| Sensibilisierung durch journalistische Standards |
| Keine spezifischen Massnahmen zu Deepfakes |

| |
|---|
| Sensibilisierung durch spezifische Ereignisse |
| Weiterbildung |
| Deepfakes erkennen/filtern |
| Journalistische Standards für Verifikation |
| Spezialisierte Abteilungen und Teams |
| Trend weg vom schnellen Journalismus |
| Kooperationen mit anderen Medienhäusern |
| Filtertechniken |
| Massnahmen, wenn Journalist:innen selbst betroffen sind |
| Spezifische Strategie/Notfallplan |
| Fälle bekannt |
| Spezifische Abteilungen |
| Schweiz eher unwahrscheinlich |
| Umgang mit Bedrohungen und Attacken |
| Antrieb für Arbeit trotz Gefahren/Widerständen |
| Grenzen der Bedrohung |
| Massnahmen bei Veröffentlichung von Deepfakes |
| Medienkompetenz stärken beim Publikum |
| Schwierigkeiten der Sensibilisierung |
| Journalistische Einbindung von Medienkompetenzförderung |
| Publikum binden und Rolle festigen |
| Sensibilisierung als Auftrag |
| (Re-)Orientierung an Standards und beruflicher Identität |

## A.4. Deepfake-Produktionssoftware

**Tabelle 25:** Liste von Software zur Deepfake-Produktion (eigene Recherche und Zusammenstellung, Stand: November 2022)

| Name of the App/Software | Part that is faked (Categorized) | Developer Name |
|---|---|---|
| Adobe Premiere | Face and voice | Adobe |
| Adobe's Project Morpheus | Face | Team: Sensei Applied Science and Machine Learning |
| AI voice actors | Voice | replicastudios |
| Avatarify | Face and voice | Avatarify ai |
| B612 Camera& Photo/ Video Editor | Face | SnowCorp company |
| CereVoice Me | Voice | CereProc |
| Clipchamp | Voice | Clipchamp |
| Copy Replace Face Photo Editor | Face | Revosoft Technologies PTY LTD |
| CoquiTTS | Voice | CoquiTTS |
| Corel VideoStudio | Face and voice | Corel |
| crazytalk | Face | Reallusion |
| Cupace – Cut and Paste Face Photo | Face | Picmax |
| DaFace | Face | BTSYS |
| Deep Nostalgia | Face | MyHeritage |
| Deep Voice | Voice | Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi |
| DeepFace | Face | Gigant Inc. |
| DeepFaceLab | Face | iperov |

| Name of the App/Software | Part that is faked (Categorized) | Developer Name |
|---|---|---|
| DeepFaceLive | Face | iperov |
| Deepfake Artbreeder Cartoonify | Face | Avatart |
| Deepfakes web β | Face | Deepfakes web |
| DeepFake-tf | Face | Not found |
| DeepNude | Full body | Deepnude |
| Deepswap | Face | Not found |
| Dfaker | Face | Not found |
| Digital voice studio | Voice | replicastudios |
| Disney's Deepfakes | Face | Jacek Naruniec (DisneyResearch\|-Studios), Leonhard Helminger (DisneyResearch\|Studios/ETH Joint PhD), Christopher Schroers (DisneyResearch\|Studios), Romann M. Weber (DisneyResearch\|Studios) |
| Duplicate Face – Swap Video, Deep Re Face Video | Face | Not found |
| dynalips | Voice | dynalips |
| Dynamixyz | Face | Take-Two Interactive Software, Inc |
| Face Changer App | Face | pei peng |
| Face Swap Booth – Face Changer | Face | Revosoft Technologies PTY LTD |
| Face Swap Live | Face | Laan Labs |
| Face Switch – Collage.Click | Face | collage.click |

| Name of the App/Software | Part that is faked (Categorized) | Developer Name |
|---|---|---|
| Face2Face | Face | Justus Thies, Michael Zollh¨ofer, Marc Stamminger, Christian Theobalt, Matthias Nießner, University of Erlangen-Nuremberg ,Max-Planck-Institute for Informatics and Stanford University |
| FaceApp | Face | FaceApp Technology Ltd. |
| faceApp | Face | FaceApp Technology Limited |
| FaceIT3 | Face | Huia – Tech Studio |
| FaceLab | Face | Lyrebird Studios |
| Facemagic | Face | Insight Technology LTD. / Deeoart Limited |
| FacePlay | Face | Not found |
| Faces | Face | Wowmaking |
| Faceswap | Face | Not found |
| FaceSwapGAN | Face | Not found |
| Fake Face | Face | Byzin Mobile Application Development |
| FakeApp | Face | Not found |
| Fakeyou | Voice | Not found |
| Figaro | Voice | Figaro Symphonly |
| Free Gradient App | Face | AI4Fun |
| GANs | Face and voice | Not found |
| Generated.photos | Face and voice | generated.photo |
| Headshot | Face | Reallusion |
| HitFilm Express | Face and voice | Fxhome |
| Iface | Face | Babel Corporation |
| Instagram DeepFake Bot | Face | Dominic Rampas |
| Ispeech | Voice | Ispeech |

| Name of the App/Software | Part that is faked (Categorized) | Developer Name |
|---|---|---|
| Jiggy | Face | Botika |
| Live Portrait | Face | D-ID's |
| Lovo.ai | Voice | Lovo.ai |
| Lyrebird | Voice | Descript |
| Mimic – AI Photo Face Animator | Face and voice | Mimicapp |
| MixBooth | Face | PiVi & Co |
| Murf | Voice | Murf ai |
| Overdub/ Descript | Voice | Descript |
| PaddlePaddle GAN | Face and voice | PaddlePaddle |
| Pix2pix GAN | Face and voice | Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros |
| Play.ht | Voice | Play.ht |
| Puzzle Deepfake | Face and voice | Linkpey |
| ReadSpeaker | Voice | readspeaker |
| Reallusion | Face | Reallusion |
| Realtime Voice Cloning | Voice | Not found |
| Reface | Face | Neocortext, Inc. |
| Reflect | Face | Neocortext, Inc. |
| Rephrase Personalized | Full body | rephrase.ai |
| Rephrase Studio | Face and voice | rephrase.ai |
| Resemble clone | Voice | Resemble AI |
| SimSwap | Face | Renwang Chen, Xuanhong Chen, Bingbing Ni,and Yanhao Ge |
| Smart portrait | Face | Adobe |
| Sonantic | Voice | Sonantic |

| Name of the App/Software | Part that is faked (Categorized) | Developer Name |
|---|---|---|
| Speaking Portrait | Face and voice | D-ID's |
| SpeakPic | Voice | Speakpic |
| Speechelo | Voice | Speechelo Review |
| Spik.ai | Voice | Spik.ai |
| StyleGAN | Face | Kim Seonghyeon |
| Synth | Face | Rosebud AI |
| Synthesia | Face and voice | Synthesia |
| Synthesys | Voice | Synthesys.io |
| tacotron2 from NVIDIA | Voice | NVIDIA |
| TakeBaker | Voice | Respeecher |
| Talkr Classic | Face and voice | Talkr, Inc. |
| Talkr Live | Face | Talkr, Inc. |
| Tokkingheads | Face | Rosebud AI |
| TTS from Mozilla | Voice | Mozilla |
| VoiceMaker | Voice | DeepZen |
| Voicera | Voice | Voicera |
| Voilà AI Artist | Face | Wemagine.AI |
| Wav2Lip | Voice | Not found |
| WaveNet | Voice | WaveNet |
| Wombo | Face | Wombo |
| Xpression | Face | AI_Innovations |
| Yanderify | Face | dunnousername |
| Zao | Face and voice | Changsha Shenduronghe Network Technology Co., Ltd. |

# A.5. Szenarien zu Deepfakes in der Politik

## A.5.1. Individuelle Ebene

### A.5.1.1. Szenario 1: Erpressung bzw. Einschüchterung eines Politikers

#### Szenarioübersicht

Deepfake-Technologien ermöglichen es, gefälschte oder irreführende Inhalte zu erstellen, die zum Zwecke der Erpressung eingesetzt werden. Dazu könnte eine Politikerin oder ein Politiker beispielsweise mit einem Deepfake-Video konfrontiert werden, das gefälschte Aussagen oder gefälschte sexuelle Praktiken enthält. Unter Androhung, den Inhalt an Medienorganisationen oder über Social Media zu verbreiten, könnte die Zielperson dazu erpresst werden, innerhalb einer Frist bestimmte Wünsche der Angreifer zu erfüllen. Mögliche Ziele der Angreiferinnen und Angreifer könnten sowohl krimineller, wirtschaftlicher als auch politischer Natur sein, um z.B. vertrauliche Informationen abzugreifen, sodass ein nationales Sicherheitsrisiko entsteht.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Der wahrscheinlichste Angriffstyp ist die Generierung belastender Inhalte mittels Fälschung privater Aussagen oder Handlungen, etwa Deepfake-pornografische Inhalte oder gefälschte Situationen, in denen Politikerinnen und Politiker vermeintlich Bestechungsgelder annehmen. Je nach Ressourcen der Angreifer kann eine Erpressung auf eine mehr oder weniger ausgetüftelte Recherche der Schwächen einer Zielperson aufbauen, um den Druck zu maximieren. Technisch gilt: Je mehr Text-, Audio- und Bildmaterial über eine Person online verfügbar ist, umso bessere Ergebnisse kann das Trainieren von Deepfake-Generatoren liefern.

Die Recherche von Medienmaterial über eine Politikerin oder einen Politiker ist unkompliziert. Einfache und glaubwürdige pornografische Deepfakes lassen sich bereits mit einem geringen Ressourcenaufwand produzieren. Annähernd lippensynchrone Deepfake-Videos, in den einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern hingegen etwas mehr Aufwand und Know-how.

### Adressatenkreis und Verbreitungsweise

Da es das übergeordnete Ziel ist, unter Androhung der Herausgabe belastender Inhalte Druck auf die Zielperson auszuüben, um diese zu bestimmten Handlungen oder Aussagen zu bewegen, wäre der Adressatenkreis in diesem Szenario auf die Zielperson beschränkt.

## A.5.1.2. Szenario 2: Rufschädigung eines Politikers

### Szenarioübersicht

In diesem Szenario setzen Angreifer bzw. Verursacher Deepfake-Inhalte zur Rufschädigung eines Politikers ein. Politiker werden mittels belastender Deepfake-Inhalte (Stimme, Bild oder Video) in Situationen gezeigt, die nie wirklich existierten, und ihnen werden Aussagen in den Mund gelegt, die von der realen Person nie getätigt worden sind. Diese Inhalte werden mittels sozialer Medien innerhalb kürzester Zeit an einen weiten Adressatenkreis gesendet. Das übergeordnete Ziel bei diesem Szenario ist die Schädigung des Rufs einer Politikerin oder eines Politikers, etwa um eine Wiederwahl zu sabotieren.

Die Rufschädigung ist eine zentrale Angriffsmethode, um über das jeweilige individuelle Ziel hinausgehende politische Ziele zu erreichen. Zudem sind Nebenwirkungen auch dann zu erwarten, wenn die Erreichung weiterer Ziele gar nicht beabsichtigt ist. So kann Rufschädigung auch zur Anstachelung von Gewalt gegen Politiker dienen. Wenn der Ruf eines Politikers beschädigt wird, er stark mit einer bestimmten politischen Partei oder staatlichen Institution verknüpft ist, kann auch deren Ruf in Mitleidenschaft gezogen werden. Eingesetzt zum richtigen Zeitpunkt, könnte auch Einfluss auf demokratische Wahlen oder politische Entscheidungsprozesse genommen werden. Ein kontinuierlicher erfolgreicher Einsatz von Deepfakes zu Rufschädigungszwecken könnte auch zu einem generellen Vertrauensverlust und damit zur Beschädigung des demokratischen Gemeinwesens führen, soziale Spannungen verschärfen, die öffentliche Sicherheit gefährden oder Einfluss auf internationale Beziehungen nehmen.

### Angriffstyp und erforderlicher Ressourcenaufwand

Ähnlich wie im ersten Szenario ist der wahrscheinlichste Angriffstyp die Generierung belastender Inhalte mittels Fälschung privater Aussagen oder Handlungen (etwa Deepfake-pornografische Inhalte oder gefälschte Situationen, in denen Politikerinnen und Politiker vermeintlich Bestechungsgelder annehmen), die sich in

Abhängigkeit von den Ressourcen eines Angreifers auf eine mehr oder weniger intensive Recherche der Schwächen einer Zielperson stützen. Auch in diesem Szenario gilt, dass Deepfake-Generatoren umso bessere Ergebnisse liefern, je mehr Text-, Audio- und Bildmaterial über eine Person online verfügbar ist.

Die Recherche von Medienmaterial über eine Politikerin oder einen Politiker ist unkompliziert. Annähernd lippensynchrone Deepfake-Videos, in denen einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern derzeit Aufwand und Know-how. Allerdings können auch Deepfakes mit niedriger Qualität zur Rufschädigung führen, etwa wenn ein gekennzeichnetes satirisches Deepfake beabsichtigter- oder unbeabsichtigterweise eine Politikerin oder einen Politiker diskreditiert.

Da die Wirksamkeit eines Angriffs zu Rufschädigungszwecken davon abhängt, dass der Deepfake-Inhalt möglichst schnell einem breiten Publikum gezeigt wird, könnten auch synthetische Social Botnets eingesetzt werden, um dies zu erreichen. Dies erfordert jedoch einen hohen Ressourcenaufwand.

#### Adressatenkreis und Verbreitungsweise

Weil zur Rufschädigung die Beeinflussung der öffentlichen Meinung erforderlich ist, richtet sich der Angriff an einen möglichst weiten Adressatenkreis.

Je nach Angriffskontext ist es aber auch denkbar, dass ein Deepfake-Inhalt zunächst an eine spezifische Zielgruppe gesendet wird. Wenn dieser Inhalt etwa zunächst in Chatgruppen anfälliger politischer Kreise kursiert, könnte gewährleistet werden, dass möglichst viele Menschen erreicht werden, die anfällig dafür wären, dem Inhalt Glauben zu schenken. Bis Opfer (oder andere Stellen mit einem Interesse an einer Richtigstellung der Fälschung) Gegenschritte einleiten, wäre die Zielgruppe bereits kontaktiert. Spezifische Bevölkerungsgruppen könnten auch mittels Micro-Targeting und der Einbettung des Deepfakes in Werbeinhalte erreicht werden.

### A.5.1.3. Szenario 3: Anstachelung zu Gewalttaten gegen einzelne Politikerinnen und Politiker

#### Szenarioübersicht

Belastende Deepfake-Inhalte können auch zur Anstachelung zu Gewalttaten gegen einzelne Politikerinnen und Politiker verwendet werden. Im Kontext einer

aufgeheizten öffentlichen Polit-Debatte könnten so beispielsweise einem verantwortlichen Politiker irreführende Worte in den Mund gelegt werden. Gewaltbereite Gruppen oder Einzelpersonen könnten dem gefälschten Inhalt Glauben schenken und zum Mittel der Gewalt greifen.

Ein tätlicher Angriff auf einzelne Politikerinnen und Politiker kann das übergeordnete Ziel eines Angreifers sein. Er kann aber auch weitere (beabsichtigte oder nicht beabsichtigte) Folgen haben: etwa die Rufschädigung der Zielperson oder deren Einschüchterung. Im Vorfeld einer Wahl könnte auch auf den Wahlprozess Einfluss genommen werden. Soziale Spannungen können geschaffen oder verschärft werden und gerade in Fällen, in denen ganze Gruppen zu Gewalttaten angestachelt werden, könnte die öffentliche Sicherheit gefährdet werden.

**Angriffstyp und erforderlicher Ressourcenaufwand**

Auch in diesem Szenario wird der Angriff vor allem auf gefälschten (privaten) Aussagen oder Handlungen beruhen. Ein solcher Deepfake-Inhalt wäre darauf ausgelegt, möglichst starke Emotionen bei den Rezipienten auszulösen. Annähernd lippensynchrone Deepfake-Videos, in denen einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern derzeit Aufwand und Know-how.

Wenn es das Ziel des Angreifers ist, nicht nur eine Einzelperson, sondern möglichst viele Menschen zu Gewalt anzustacheln, ist mit dem Einsatz synthetischer Social Botnets zu rechnen. Dies würde einen hohen Ressourceneinsatz erfordern.

**Adressatenkreis und Verbreitungsweise**

In Abhängigkeit davon, ob eine Einzelperson oder eine grössere Anzahl an Menschen zu Gewalt angestachelt werden sollen, kann es sich um einen engen oder weiteren Adressatenkreis handeln.

Wie in Szenario 2 ist es möglich, dass ein Deepfake-Inhalt zunächst an eine spezifische Audienz gesendet wird. Wenn der Inhalt etwa zunächst in Chatgruppen anfälliger politischer Kreise kursiert, könnte gewährleistet werden, dass ihm möglichst viele potenziell anfällige Menschen Glauben schenken, bevor die Opfer (oder andere Stellen) Gegenschritte einleiten könnten. Spezifische Bevölkerungsgruppen könnten auch mittels Micro-Targeting und der Einbettung des Deepfakes in Werbeinhalte erreicht werden.

## A.5.2. Institutionelle Ebene

### A.5.2.1. Szenario 4: Rufschädigung einer Partei oder (staatlicher) Institution

#### Szenarioübersicht

Deepfake-Inhalte, die zur Rufschädigung eingesetzt werden, können auch zum Ziel haben, das Vertrauen in eine Partei oder staatliche Institution zu untergraben (Verwaltung innovativ 2023). In diesem Szenario dient ein Deepfake-Inhalt nicht primär der Diskreditierung einer Person bzw. eines Politikers. Die Beschädigung des Rufs einer Politikerin oder eines Politikers ist vielmehr ein Mittel, um das Vertrauen in eine Partei und/oder staatliche Institution zu senken. Einer Amtsinhaberin einer politischen Institution könnten etwa Aussagen in den Mund gelegt werden, denen zufolge sie in ihrem Amt im Auftrag ihrer Partei Manipulationen vornimmt.

Wie im Falle der individuellen Rufschädigung können auch in diesem Szenario weitere politische Ziele verfolgt werden bzw. Nebenwirkungen auftreten: Anstachelung zur Gewalt, Einschüchterung, Beeinflussung einer Wahl und von politischen Entscheidungsprozessen, die Beschädigung des demokratischen Wesens, die Verschärfung sozialer Spannungen, die Gefährdung der öffentlichen Sicherheit und die Beeinflussung internationaler Beziehungen.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Grundlage für dieses Szenario ist die Generierung belastender Inhalte mittels Fälschung privater Aussagen oder Handlungen.

Auch in diesem Szenario kann auf synthetische Social Botnets zurückgegriffen werden, um die Verbreitung der Inhalte zu beschleunigen und zu verstärken. Dies impliziert einen hohen Ressourcenaufwand.

#### Adressatenkreis und Verbreitungsweise

Der genaue Adressatenkreis hängt mit dem konkreten Ziel des Angriffs zusammen: Wenn es das Ziel ist, ein Angriffsziel (Partei oder politische Institution) in den Augen möglichst vieler Menschen zu diskreditieren, würde der Deepfake an einen möglichst weiten Adressatenkreis verteilt werden.

Wenn es hingegen stärker darum geht, das Vertrauen in eine Partei oder politische Institution bei einer bestimmten Gruppe zu verringern, würde der Deepfake-Inhalt zunächst in dieser Gruppe verbreitet, um eine Entdeckung möglichst lange zu vermeiden. Spezifische Bevölkerungsgruppen könnten mittels teilöffentlicher Kommunikationskanäle wie Chatgruppen sowie mit der Einbettung des Deepfakes in Werbeinhalte unter Einsatz von Micro-Targeting erreicht werden.

### A.5.2.2. Szenario 5: Abgreifen von vertraulichen Informationen

#### Szenarioübersicht

In diesem Szenario dienen Deepfakes dazu, an vertrauliche Informationen zu gelangen, indem entweder technische Schutzmassnahmen überwunden werden oder *Social Engineering* gegen Personen zum Einsatz kommt. Indem mittels Deepfake-Technologie Identitätsdiebstahl begangen wird, können z.B. biometrische Authentifizierungssysteme mit Gesichts- und/oder Stimmerkennung überlistet werden. Ebenso können mittels Deepfakes Personen getäuscht und zur Durchführung von Handlungen, etwa der Herausgabe von Passwörtern, manipuliert werden. Während Social Engineering im finanziellen Bereich primär zur wirtschaftlichen Gewinnerzielung eingesetzt wird, ist zu erwarten, dass es in der Politik vor allem zum Erbeuten von vertraulichen Informationen genutzt wird.[290]

Hierarchische Strukturen, wie sie in der Verwaltung aus Effizienzgründen üblich sind (Nullmeier 2010), können bei einer Social-Engineering-Attacke als verstärkender Faktor wirken, wenn etwa Mitarbeitende ihre Zweifel an der Authentizität unausgesprochen lassen oder Zweifel erst gar nicht zulassen.

Dieses Szenario ist mit anderen Szenarien verbunden: Abgegriffene vertrauliche Informationen können zur Beeinflussung einer Wahl und von politischen Entscheidungsprozessen genutzt werden. Im Falle des Bekanntwerdens des Informationsdiebstahls kann das Vertrauen in das politische System sinken und so eine Beschädigung des demokratischen Gemeinwesens die Folge sein. Wenn beispielsweise nachrichtendienstliche Informationen abgegriffen werden, könnten diese zur Verschärfung sozialer Spannungen und zur Gefährdung der öffentlichen Sicherheit missbraucht werden. Wenn vertrauliche staatliche Informa-

[290] Der Giffey-Klitschko-Fall diente zwar vordergründig satirischen Zwecken des russischen Komikerduos. Bei einem vergleichbaren Angriff könnten jedoch durchaus vertrauliche Informationen erbeutet werden.

tionen erbeutet werden, können schliesslich auch internationale Beziehungen betroffen sein.

### Angriffstyp und erforderlicher Ressourcenaufwand

Bei diesem Szenario ist Identitätsdiebstahl die Grundlage für zwei Angriffstypen. Indem das Aussehen und die Stimme z.B. eines Mitarbeitenden imitiert werden, könnten Angreifer erstens Zugang zu vertraulichen Informationen erhalten, bei denen sich eine Person mittels Stimm- oder Gesichtsbiometrie authentifizieren muss. Dies könnte eine physische Zugangskontrolle zu einem Gebäude oder Raum sein, aber auch der Zugang zu einem digitalen Dienst oder Endgerät wie einem Computer oder Smartphone. Der Zugriff auf (biometrische) Daten über eine Zielperson und die Nutzung derartiger Daten zur Erstellung eines Deepfakes, das geeignet wäre, um gesichtsbiometrische Zugangskontrollen zu überwinden, erfordern einen sehr hohen Ressourcenaufwand. Etwas einfacher kann die Überwindung stimmbiometrischer Zugangskontrollen ausfallen.

Indem Aussehen und Stimme einer Person imitiert werden, könnten ausserdem Zielpersonen getäuscht und zur Durchführung bestimmter Handlungen manipuliert werden, beispielsweise indem ein imitierter Vorgesetzter die Herausgabe vertraulicher Informationen anordnet. Die Imitation der Stimme von Vorgesetzten ist mit relativ geringem Aufwand möglich. Die Imitation des Aussehens und der Stimme in Echtzeit erfordern einen deutlich höheren Ressourcenaufwand.

### Adressatenkreis und Verbreitungsweise

Weil ein solcher Angriff die Hilfsbereitschaft, das Vertrauen, die Angst oder den Respekt vor dem Gegenüber ausnutzt, ist der Adressatenkreis auf einzelne Mitarbeitende begrenzt, um einer Entdeckung zu entgehen. Sowohl einzelne Mitarbeitende als auch ganze Teams und Mitarbeiterstäbe können je nach Qualität des Deepfakes getäuscht werden.

## A.5.3. Gesellschaftliche Ebene

### A.5.3.1. Szenario 6: Beeinflussung einer demokratischen Wahl

#### Szenarioübersicht

Das in der Deepfake-Literatur meistdiskutierte Szenario ist die Beeinflussung von demokratischen Wahlen mittels Deepfakes. Typischerweise werden in diesem Szenario einzelnen Politikerinnen und Politikern vor Wahlen falsche Aussagen in den Mund gelegt, um einen entscheidenden Einfluss auf den Ausgang einer Wahl, einer Volksinitiative oder eines Referendums zu nehmen.

Je nach politischer Lage können unterschiedliche Gruppen im Fokus von Kampagnen zur Wahlbeeinflussung stehen: darunter die Gruppe der Unentschlossenen und Wechselwähler. Politische Parteien und Gruppen können ausserdem die eigene Wählerschaft ansprechen oder auch die Wähler der politischen Konkurrenz.

Denkbar ist es auch, dass im Vorfeld einer Volksinitiative Deepfakes dazu genutzt werden, anfällige Bevölkerungsteile zur Abgabe von Unterschriften zu manipulieren.

Wie die Diskussion in Kapitel 4.4.2 gezeigt hat, kann die nachgewiesene Beeinflussung einer Wahl dazu führen, dass die Ergebnisse nicht anerkannt werden und eine Wahl wiederholt werden muss.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Bekannte Vorfälle zeigen, dass Versuche der Einflussnahme auf demokratische Wahlen v.a. auf Deepfake-Inhalte mit gefälschten privaten Aussagen oder Handlungen basieren. Konkret könnten einer Zielperson beispielsweise rufschädigende Aussagen in den Mund gelegt werden, um diese in der Wahrnehmung der Wahlbevölkerung zu diskreditieren. Auch in der Literatur zu den Herausforderungen von Deepfakes steht die Gefahr dieser Art von Deepfakes im Vordergrund. Die Recherche von Medienmaterial über eine Politikerin oder einen Politiker ist unkompliziert. Annähernd lippensynchrone Deepfake-Videos, in denen einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern derzeit Aufwand und Know-how.

Unter hohem Ressourceneinsatz können synthetische Social Botnets genutzt werden, um mit dem Deepfake einen möglichst grossen Adressatenkreis zu erreichen.

#### Adressatenkreis und Verbreitungsweise

Auch bei diesem Szenario hängt der Adressatenkreis vom konkreten Angriffsziel ab: Die Beeinflussung der Gruppe der Unentschlossenen und Wechselwähler verspricht vor allem dann Erfolg, wenn bereits ein knapper Wahlausgang erwartet wird, der mittels Deepfake-Einsatzes in eine gewünschte Richtung gekippt werden soll. Entscheidend für den Erfolg eines derartigen Angriffs ist die vorherige Informationssammlung über Bevölkerungsgruppen, die als anfällig für entsprechende Inhalte gelten. Diese könnten mittels Micro-Targeting und der Einbettung des Deepfakes in Werbeinhalte erreicht werden oder durch das Posten der Inhalte in einschlägigen Foren, Chatgruppen usw.

Wenn es das Ziel ist, einen radikalen Bevölkerungsteil weiter zu radikalisieren, um ihn somit zu einer intensivierten Unterstützung des Wahlkampfes oder bestimmter politischer Positionen, Akteure usw. zu bewegen, würden eher private oder halböffentliche Kommunikationskanäle verwendet werden. Auch in diesem Fall ist ein entscheidender vorbereitender Schritt die Informationssammlung über den Bevölkerungsteil, der als anfällig für entsprechende Inhalte gilt. Auf sozialen Netzwerken könnten diese Menschen mittels Micro-Targeting und (Chat-) Gruppen usw. erreicht werden.

Schliesslich könnte sich ein Deepfake auch an weite Teile der Bevölkerung richten. Gerade im unmittelbaren Vorfeld einer Wahl könnte ein solcher, grossangelegter Deepfake-Einsatz unter zusätzlichem Einsatz von synthetischen Social Bots das Ziel verfolgen, möglichst viele Menschen zu erreichen, um ihre Wahlentscheidung in eine bestimmte Richtung zu manipulieren.

### A.5.3.2. Szenario 7: Beeinflussung von politischen Entscheidungsprozessen

#### Szenarioübersicht

Deepfakes ermöglichen nicht nur die Beeinflussung von Wahlen. Auch politische Entscheidungsprozesse in Parlamenten, Verwaltungen usw. können mittels Deepfakes beeinflusst werden. Neben der Erpressung, Einschüchterung oder Diskreditierung wichtiger Entscheidungspersonen mittels Deepfakes stellt

sich eine neue Herausforderung insb. durch digitales Astroturfing, also der Vortäuschung einer spontanen zivilgesellschaftlichen Bewegung. Die mittels neuer Deepfake-Technologien ermöglichte Erschaffung authentisch wirkender Bots kann auch zum massenhaften Lobbying für oder gegen bestimmte Gesetzesvorhaben verwendet werden. Erfolgreich könnte digitales Astroturfing vor allem dann sein, wenn es während der Agenda Setting- bzw. Formulierungsphase eines Gesetzes stattfindet, um beispielsweise während einer öffentlichen Konsultation massenhaft Einreichungen einzusenden, die grundsätzlich in eine Richtung gehen, aber doch so authentisch formuliert sind, dass unerkannt bleibt, dass sie von Bots erstellt wurden. Eine Beeinflussung in späteren Phasen eines Entscheidungsprozesses wäre möglich, indem synthetische Social Botnets Kampagnen auf sozialen Medienplattformen organisieren.

**Angriffstyp und erforderlicher Ressourcenaufwand**

Sofern digitales Astroturfing angewandt wird, würden vor allem synthetische Social Bots genutzt werden, um den Eindruck einer Massenbewegung zu erzeugen. Konkrete Form annehmen würde ein solcher Deepfake-Einsatz z.B. in der Weise, dass massenhaft authentisch wirkende Meinungsäusserungen für oder gegen bestimmte politische Entscheidungen oder Pläne im Rahmen einer politischen Konsultation eingereicht werden. Einflussnahme ist auch mittels Social-Media-Kampagnen denkbar, wenn synthetische Social Botnets diese in koordinierter Weise initiieren und nähren. Unter zusätzlichen Druck könnten Entscheidungstragende gesetzt werden, wenn es die Kampagne schafft, natürliche Personen z.B. zu öffentlichen Demonstrationen zu motivieren. Diese Methoden erfordern allesamt einen sehr hohen Ressourcenaufwand.

Eine Beeinflussung politischer Entscheidungsprozesse wäre allerdings auch mittels Fälschung privater Aussagen oder Handlungen möglich: Wichtige Entscheidungsträger könnten mittels diskreditierender Deepfakes oder durch Deepfake-basierte Erpressung aus dem Amt gedrängt werden, wenn z.B. in Aussicht steht, dass deren Nachfolger willkommenere Positionen vertreten. Wenn belastende Deepfakes zur Einschüchterung verwendet werden, könnte selbst das Entscheidungsverhalten von Parlamentariern und Amtsträgern manipuliert werden. Annähernd lippensynchrone Deepfake-Videos, in denen einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern weniger Aufwand und Know-how als die o.g. Methoden zu digitalem Astroturfing und Social Botnets.

#### Adressatenkreis und Verbreitungsweise

Wenn es das Ziel wäre, politische Entscheidungstragende in frühen Gesetzesphasen, etwa während der Formulierung eines Vorschlags zu beeinflussen, würde eine Einflussnahme auf den engen Adressatenkreis der entsprechenden Entscheidungstragenden fokussieren. Typische Kommunikationskanäle wären entweder Einreichungen während formeller Konsultationen oder direkte E-Mail-Kommunikation.

Wenn die Einflussnahme in einer späteren Phase des Gesetzgebungsprozesses erfolgt, könnte als Hebel zur Beeinflussung der Entscheidungstragenden die breite Öffentlichkeit genutzt werden. In diesem Fall würden soziale Medienplattformen, Chatgruppen usw. als Mittel verwendet werden, um eine möglichst breite öffentliche Aufmerksamkeit auf ein Thema zu lenken, indem Emotionen zu dem Thema erzeugt werden.

Ein weiter Adressatenkreis wäre auch dann relevant, wenn wichtige Entscheidungspersonen mittels diskreditierender Deepfakes aus dem Amt gedrängt werden sollen.

Im Falle der Beeinflussung der individuellen Entscheidung von Amtsträgern oder Parlamentariern würden diese Personen mittels erpresserischer und einschüchternder Deepfakes direkt kontaktiert werden.

### A.5.3.3. Szenario 8: Verschärfung sozialer Spannungen

#### Szenarioübersicht

Deepfakes können dazu eingesetzt werden, soziale Spannungen zu verschärfen. Wenn Politikern Worte in den Mund gelegt werden, mit denen sie Teile der Bevölkerung vermeintlich beleidigen, kann dies in einen Vertrauensverlust dieses Bevölkerungsteils und zur Verschärfung von Spannungen münden. Weil zudem Desinformation und irreführende Inhalte selbst nach mühevoller Richtigstellung bei einem Teil der Betrachter eine anhaltende Wirkung hinterlassen, ist davon auszugehen, dass soziale Spannungen nicht von einem einzelnen Deepfake ausgelöst oder verschärft werden, sondern erst mit der kontinuierlichen Verbreitung von Deepfakes allmählich stärker werden.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Desinformationsakteure nutzen verschiedene Mittel zur Verschärfung sozialer Spannungen, darunter die Unterstützung politischer Parteien und zivilgesell-

schaftlicher Kräfte, die ihren Zielen dienlich sind, oder die Manipulation der öffentlichen Meinung mittels Micro-Targeting auf sozialen Onlinemedienplattformen. Deepfakes könnten sich in das Werkzeugrepertoire dieser Kräfte einreihen.

Ähnlich wie bei vielen der anderen Szenarien könnten auch in diesem Szenario gefälschte private Aussagen oder Handlungen verwendet werden, um möglichst starke Emotionen bei den Adressaten auszulösen. Annähernd lippensynchrone Deepfake-Videos, in denen einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern derzeit Aufwand und Know-how.

Wenn es das Ziel wäre, nicht nur einzelne Bevölkerungsteile zu erreichen, könnten unter einem hohen Ressourceneinsatz auch synthetische Social Bots zum Einsatz kommen, um einen möglichst breitflächigen Effekt zu erreichen.

**Adressatenkreis und Verbreitungsweise**

Wenn soziale Spannungen mit Absicht erschaffen oder verschärft werden, können zwei Adressatenkreise im Fokus stehen. Zum einen bereits radikalisierte oder für irreführende Inhalte anfällige Bevölkerungsteile, deren weitere Radikalisierung als Katalysator für soziale Spannungen wirken könnte. Diese Bevölkerungsteile würden über eher private und teilöffentliche Kommunikationskanäle über die Einbettung des Deepfakes in Werbeinhalte und die Aussendung dieser mittels Micro-Targeting adressiert. Sofern es sich dabei um Personen handelt, die sich bereits in Echokammern befinden, wäre die Wahrscheinlichkeit erhöht, dass sie von etwaigen Richtigstellungen nicht oder nur spät erreicht werden oder diese bewusst oder unbewusst ignorieren.

Zum anderen könnte die breite Bevölkerung im Fokus derartiger Bestrebungen stehen. Hierbei sind wiederum zwei Varianten denkbar:

In der ersten Variante könnte die breite Masse mit dem einfachen Ziel adressiert werden, langfristig «Informationsverschmutzung» zu betreiben. Richtigstellungen würden viele Menschen zeitnah erreichen und verhindern, dass irreführende Deepfake-Inhalte einen grossen unmittelbaren Effekt nach sich ziehen. Weil es als schwierig gilt, mit Richtigstellungen alle Betrachter eines irreführenden Inhalts zu erreichen, und desinformative Inhalte häufig trotzdem einen bleibenden Effekt hinterlassen, könnte es das Ziel sein, mittels der regelmässigen Aussendung derartiger Deepfakes an die breite Masse eine allmähliche Verunsicherung über den Wahrheitsgehalt von Nachrichten und Informationen in der Bevölkerung zu bewirken.

In der zweiten Variante könnte ein qualitativ sehr hochwertiger Deepfake in einer ohnehin aufgeladenen Situation in Umlauf gebracht werden, um die Öffentlichkeit weiter zu polarisieren. Aufgrund der in einer solchen Situation aufkommenden Emotionen auf beiden Seiten der Debatte könnte die Wirkung von etwaigen Richtigstellungen weiter vermindert werden und der Deepfake eine besonders grosse Wirkung entfalten.

### A.5.3.4. Szenario 9: Beschädigung der Demokratie

#### Szenarioübersicht

Eine weitere vieldiskutierte Gefahr von Deepfakes ist deren Potenzial, die Demokratie zu beschädigen. In diesem Szenario geht es weniger um die unmittelbaren Folgen einzelner Deepfakes. Die Erosion der Demokratie ist vielmehr als eine schleichende (Spät-)Folge der Gesamtheit unterschiedlicher Nutzungen von Deepfakes zu verstehen.

Mehrere miteinander verbundene Faktoren gelten in Debatten rund um die Schädigung der Demokratie als relevant (OECD 2022): geringe Beteiligungsraten an Wahlen und Referenden, die Polarisierung der Gesellschaft bzw. von öffentlichen Debatten sowie abnehmendes bzw. geringes Vertrauen in das politische System und in Nachrichtenmedien. In allen Bereichen schneidet die Schweiz im Vergleich zu vielen anderen Demokratien tendenziell gut ab: Die Beteiligungsraten sind zuletzt wieder gestiegen (Bundesamt für Statistik 2023). Von einer polarisierten Öffentlichkeit ist kaum die Rede (Tribelhorn 2022). Das Vertrauen in das politische System ist in den vergangenen Jahren kontinuierlich gestiegen (Bundesamt für Statistik 2022). Das Medienvertrauen ist in absoluten Werten zwar nicht besonders hoch, im internationalen Vergleich allerdings durchaus – und es bleibt seit Jahren konstant (Newman 2022: 107).

Trotzdem wird Deepfakes zugetraut, diese relative Stabilität zu beeinträchtigen. Diese Perspektive spiegelt sich auch in den Umfrageergebnissen mit Schweizer Stände- und Nationalräten wider (vgl. Abschnitt 6.2). Im Rahmen dieses Szenarios sollen nun die Möglichkeiten der Beschädigung der Demokratie in der Schweiz diskutiert werden.

Die in Szenario 7 diskutierte Wahlbeeinflussung mittels Deepfakes ist ein Mittel, das zur unmittelbaren Senkung der Wahlbeteiligung eingesetzt werden könnte. Wie in Szenario 8 diskutiert, könnten Deepfakes auf verschiedene Weisen eingesetzt werden, um soziale Spannungen zu verschärfen. Mittels Schädigung

des Rufes von Politikern und von politischen Institutionen könnte das Vertrauen in das politische System untergraben werden. Durch allgegenwärtige Deepfake-induzierte «Informationsverschmutzung» könnten schliesslich auch das Vertrauen in Informationsflüsse im Allgemeinen und das Medienvertrauen erodieren (siehe auch Kapitel 5).

KI-generierte Bilder sind längst in der Lage, Millionen von Menschen in die Irre zu führen (Keller 2023). Auch können Politiker selbst durch einen Deepfake getäuscht werden, sodass sie diesen verbreiten oder anderweitig darauf reagieren. Sofern es Menschen gibt, die einem irreführenden Inhalt Glauben schenken, ist es für die Beschädigung der Debattenkultur zudem unerheblich, ob Fahrlässigkeit oder der Wunsch nach satirischer Unterhaltung ausschlaggebend für das Teilen eines Inhalts waren (Daloz 2021).

Die Beschädigung der Demokratie kann schliesslich auch durch die sog. «Dividende des Lügners» (Citron/Chesney 2019) eintreten: Denn wenn Wahres und Unwahrheiten zunehmend schwieriger auseinandergehalten werden können, fällt es jenen, die Unwahrheiten oder irreführende Inhalte verbreiten, einfacher, beim Publikum Gehör zu finden. Am Ende stünden nicht mehr Wahrheit und Lüge gegenüber, sondern nur noch «alternative Wahrheiten».

Daneben besteht die Möglichkeit, dass die Beschädigung der Demokratie infolge der unbeabsichtigten vermehrten Nutzung von Deepfakes allmählich erfolgt. Wie in mehreren vorherigen Szenarien diskutiert, kann das gutgemeinte Teilen von Deepfake-Inhalten seitens der breiten Bevölkerung dennoch zu nicht intendierten Effekten wie der Rufschädigung von Politikern und politischen Institutionen, der Beeinflussung von Wahlen und der Verschärfung sozialer Spannungen führen, die ihrerseits wiederum Folgen für die Demokratie nach sich ziehen können. Solche nicht intendierte Effekte können zudem auch das Ergebnis des Handelns von politischen oder zivilgesellschaftlichen Akteuren sein, wie der Fall des Deepfakes in der belgischen Klimapolitik oder der Rückgriff auf KI-generierte Bilder seitens Amnesty International, die in der Kritik standen, zur Informationsverschmutzung beizutragen, verdeutlichen.

**Angriffstyp und erforderlicher Ressourcenaufwand**

Weil das Szenario der Beschädigung der Demokratie eng mit vielen weiteren Szenarien zusammenhängt, kommen verschiedene der bereits in den o.g. Szenarien diskutierten Angriffstypen infrage. Die häufigsten und wahrscheinlichsten Angriffstypen dabei sind allerdings die Fälschung von (privaten) Aussagen und

Handlungen und der Rückgriff auf synthetische Social Botnets, um Deepfake-Inhalte effektiver zu verbreiten.

#### Adressatenkreis und Verbreitungsweise

Zum Zwecke der Beschädigung der Demokratie eingesetzte Deepfake-Inhalte werden voraussichtlich einen weiten Adressatenkreis haben, um eine möglichst grosse Wirkung auf die breite Masse der Bevölkerung auszuüben.

### A.5.3.5. Szenario 10: Gefährdung der öffentlichen Sicherheit

#### Szenarioübersicht

Die Gefährdung der öffentlichen Sicherheit kann ebenfalls mittels verschiedener Deepfake-Angriffe das Ziel von Angreifern sein. Dazu könnte ein «emotionalisierender» Deepfake eine ohnehin aufgeheizte politische Situation weiter verschärfen. Wie der Angriff auf das US-Kapitol Anfang 2022 eindrucksvoll gezeigt hat, können Massen mittels einer Lüge (in dem Fall, des vermeintlichen Wahlbetrugs) zu gewaltsamen Aktionen verleitet werden, um ihre politischen Ziele zu erreichen. Ein Deepfake (in dem bspw. der vermeintliche Wahlbetrug gezeigt wird oder vermeintliche Drahtzieher über den Wahlbetrug sprechen), das zum richtigen Zeitpunkt verbreitet wird, könnte der notwendige Katalysator für Gewalt sein. Soziale Unruhen könnten jedoch auch in beliebigen anderen Kontexten mittels Deepfakes befeuert werden: etwa ein Deepfake, der vermeintlich übermässige Polizeigewalt darstellt oder Demonstranten bei einem vermeintlichen gewalttätigen Angriff auf Polizistinnen und Polizisten zeigt.

Darüber hinaus kann die öffentliche oder nationale Sicherheit auch dadurch gefährdet werden, dass unter Einsatz von Deepfakes vertrauliche Informationen abgegriffen werden (vgl. auch Szenario 5). Dies umfasst sowohl den Zugriff auf staatliche (z.B. nachrichtendienstliche) Daten, deren Missbrauch geeignet ist, die öffentliche Sicherheit zu gefährden, als auch auf sonstige kritische Daten. Hierbei wäre z.B. an Steuerungsdaten kritischer Infrastruktur zu denken.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Gefährdungen der öffentlichen Sicherheit können auf gefälschten privaten Aussagen oder Handlungen beruhen, die an bestimmte, gewaltaffine Bevölkerungsteile kommuniziert werden, um eine entsprechende Reaktion zu bewirken. Das

Micro-Targeting derartiger Inhalte erfordert viel Know-how und einen recht hohen Ressourcenaufwand.

Insbesondere in Situationen, in denen eine kontroverse gesellschaftliche Debatte herrscht, könnten entsprechende Inhalte mittels synthetischer Social Bots auch an die breite Bevölkerung kommuniziert werden.

Die öffentliche und nationale Sicherheit kann auch dadurch gefährdet werden, dass unter hohem Ressourcenaufwand stimm- und gesichtsbiometrische Authentifizierungssysteme mittels Deepfakes und Social Engineering überwunden werden, um nicht autorisierten Zugang zu vertraulichen Informationen zu erhalten.

**Adressatenkreis und Verbreitungsweise**

Wenn die Gefährdung der öffentlichen Sicherheit mittels Mobilisierung von gewaltbereiten Gruppen im Fokus steht, würde sich der Adressatenkreis auf die jeweiligen Bevölkerungsteile beziehen. Diese würden vor allem über private und teilöffentliche Kanäle kontaktiert, um eine maximale Wirkung zu erzielen, bevor Massnahmen wie Richtigstellungen oder Löschung der jeweiligen Inhalte ergriffen oder sonstige (polizeiliche usw.) Abwehrmassnahmen eingeleitet werden können. Ein Element in der Adressierung dieser Bevölkerungsteile könnte die Einbettung des Deepfakes in Werbeinhalte und deren Aussendung mittels Micro-Targeting sein.

Wenn die Mobilisierung sich auf die gesamte Gesellschaft bezieht, würden öffentliche Kanäle, insb. soziale Medienplattformen genutzt werden.

Wenn die Gefährdung der öffentlichen Sicherheit über die Erschleichung von Zugang zu vertraulichen Informationen unter Ausnutzung der Hilfsbereitschaft, des Vertrauens, der Angst oder des Respekts vor dem Gegenüber erfolgt, wäre der Adressatenkreis so eng wie möglich, um einer Entdeckung zu entgehen. Eng bedeutet allerdings nicht zwingend, dass ausschliesslich Einzelpersonen getäuscht werden. Auch Mitarbeitende von z.B. Politikerinnen und Politikern könnten Opfer der Täuschung sein.

### A.5.3.6. Szenario 11: Beeinflussung der internationalen Beziehungen

#### Szenarioübersicht

Schliesslich können Deepfakes auch zur Beeinflussung der internationalen Beziehungen verwendet werden, etwa indem Politikerinnen und Politikern Worte mit internationaler Brisanz in den Mund gelegt werden. Audiovisuelle Deepfakes und manipulierte Nachrichtenbilder sind gleichermassen dazu geeignet, internationale Beziehungen zu destabilisieren.

Zudem könnte die Beeinflussung der internationalen Beziehungen nicht nur mittels eines Deepfakes selbst erfolgen, sondern auch unter Rückgriff auf vertrauliche staatliche Informationen, die zuvor unter Einsatz von Deepfakes abgegriffen wurden, wie sie in Szenario 5 diskutiert wurden.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Wenn der Deepfake-Angriff auf der Irreführung von Öffentlichkeit und Politik beruht, müssen gefälschte Aussagen oder Handlungen einer möglichst breiten Öffentlichkeit bekannt gemacht werden. Zur Steigerung der Wirkung könnten synthetische Social Bots eingesetzt werden.

Wenn der Deepfake-Angriff auf der Erbeutung von vertraulichen staatlichen Informationen beruht, würde der Angriff zunächst auf der Überwindung von stimm- und gesichtsbiometrischen Sicherheitsmassnahmen mittels Deepfakes und Social Engineering basieren.

#### Adressatenkreis und Verbreitungsweise

Falls es den Angreifern um die Irreführung der Öffentlichkeit und Politik geht, könnten Deepfakes entweder direkt an die zu manipulierenden Adressaten (ein ausländisches Ministerium, ein ausländischer Politiker usw.) versendet werden oder über soziale Medienplattformen und Massenmedien an eine möglichst breite Öffentlichkeit in der Schweiz oder der Welt kommuniziert werden.

Im Falle der Erbeutung vertraulicher staatlicher Informationen mittels Social Engineering wäre der Adressatenkreis auf die jeweiligen Zielpersonen beschränkt. Die Überwindung von stimm- und gesichtsbiometrischen Sicherheitsmassnahmen hat keinen menschlichen Adressatenkreis, weil ausschliesslich Computer im Fokus stünden.

# A.6. Szenarien zu Deepfakes in der Wirtschaft

## A.6.1. Individuelle Ebene

### A.6.1.1. Szenario 1: Identitätsdiebstahl bzw. -betrug

#### Szenarioübersicht

Individuen könnten das Ziel von Deepfake-basiertem Identitätsdiebstahl werden. Die breite Verfügbarkeit von Audio- und Bildinhalten im Internet, die viele Menschen entweder freiwillig öffentlich teilen oder die über Datenlecks verfügbar werden, eröffnet diese Möglichkeit. Zu erwarten ist, dass bekannte Formen des Identitätsdiebstahls, wie der Dokumentenbetrug und die Imitierung der Stimme, um Deepfake-basierte Angriffstypen ergänzt werden.

Einerseits können solche Angriffe dazu genutzt werden, um unternehmensseitige Sicherheitsmassnahmen zu überwinden, um etwa Zugriff auf das Bankkonto einer Person zu erhalten.

Andererseits können Individuen geschädigt werden, indem sie z.B. Opfer eines Deepfakes werden, in dem sich der Angreifer für jemand anderes ausgibt. Dieser Angriff kann ganz unterschiedliche Formen annehmen, basiert aber in der Regel auf der Ausnutzung des Vertrauens der Zielperson(en) gegenüber der imitierten Person. Mitte 2023 waren Kryptobetrüger etwa in der Lage, Kryptowährungen im Wert von mind. 35 000 US-Dollar zu erbeuten. Dazu erstellten sie ein Deepfake-Video von Elon Musk, der darin vermeintlich verspricht, auf ein Konto Kryptowährung zu überweisen, um die doppelte Menge zurückzuerhalten (Lampert 2023).

Ein anderes Szenario in diesem Zusammenhang ist die Deepfake-Variante des Enkeltricks, bei dem sich Angreifer mittels eines Deepfakes für den Enkel in Not (oder dessen Freund) ausgeben, um besorgte Grosseltern zur Überweisung einer Geldsumme zu bewegen (Ludewig 2023).

#### Angriffstyp und erforderlicher Ressourcenaufwand

Mittels Gesichtsmorphing, Gesichtsaustausch, Gesichtsgenerierung und dem Klonen der Stimme lassen sich unternehmensseitige Sicherheitsmassnahmen überwinden, um Know-Your-Customer-Mechanismen zu umgehen. Beispielsweise konnte Anfang 2023 ein stimmerkennungsbasiertes Authentifizierungs-

system einer Bank mittels einer mit wenigen Ressourcen erstellten geklonten Stimme umgangen werden (Wiegand 2023).

Gesichtsaustausch und das Klonen der Stimme können auch für einen Social-Engineering-Angriff verwendet werden, um eine Zielperson beispielsweise zur Überweisung eines Geldbetrags oder – wie im Falle des Enkeltricks üblich – der Übergabe von Geld an Dritte zu bewegen. Die Fälschung der Stimme ist bereits heute möglich. Wenn es möglich bzw. einfacher zugänglich wird, Echtzeit-Deepfake-Videos zu produzieren, könnten Zielpersonen zur Steigerung der Glaubwürdigkeit per Videoanruf kontaktiert werden.

#### Adressatenkreis und Verbreitungsweise

Im Falle der Umgehung von stimm- und gesichtsbiometrischen Sicherheitsmassnahmen gibt es im engeren Sinne keinen menschlichen Adressatenkreis, weil ausschliesslich Computer im Fokus stünden. Im weiteren Sinne ist allerdings das jeweilige Unternehmen für die Sicherheit des verwendeten Systems verantwortlich.

Im Falle des Identitätsbetrugs mittels eines Social-Engineering-Angriffs ist der Adressatenkreis auf die jeweilige Zielperson beschränkt. Die Zielgruppe eines Social-Engineering-Angriffs im Falle eines Deepfake-basierten Enkeltricks steigert die Gefahr eines solchen Angriffs noch weiter: Weil es sich dabei in der Regel um ältere Personen handelt, besteht die Gefahr, dass beispielsweise Deepfake-Artefakte, die Verdacht erregen könnten, nicht als solche wahrgenommen werden oder auf eine schlechte Internetanbindung zurückgeführt werden.

### A.6.2. Organisationsebene

#### A.6.2.1. Szenario 2: Identitätsbetrug gegenüber Unternehmen

#### Szenarioübersicht

Mittels Deepfakes ist Identitätsbetrug nicht nur zulasten natürlicher Personen möglich, sondern auch gegenüber Unternehmen.

Klassische Formen der Dokumentenfälschung (z.B. von Ausweisen und Urkunden) werden seit Längerem im Finanzbereich zur Erbeutung von Geldsummen verwendet, indem ein Individuum zahlreiche gefälschte Ausweise dazu verwendet, bei vielen verschiedenen Finanzinstitutionen Konten anzulegen bzw. Kredi-

te zu erhalten. Im Laufe der Jahre wurden verschiedene Schutzmassnahmen, wie stimm- und gesichtsbiometrische Know-Your-Customer-Mechanismen, seitens der betroffenen Unternehmen eingeführt, um derartige Betrugsversuche abzuwehren.

Mittels Deepfake-Technologie sind inzwischen neue Formen der Dokumentenfälschung und der Umgehung von stimm- und gesichtsbiometrischen Sicherheitsmassnahmen möglich geworden. Mittels Gesichtsgeneratoren können Angreifer z.B. beliebig viele synthetische Gesichter bzw. Identitäten erzeugen, um Konten auf deren Namen anzulegen. Mittels Deepfake-Videotechnologie können dann gesichtsbiometrische Sicherheitsmassnahmen in Echtzeit überwunden werden.

Durch den Rückgriff auf Gesichtsmorphing-Technologie ist es zudem möglich, solche Fotos für Dokumente zu erzeugen, die noch eine ausreichend hohe Übereinstimmung mit dem ursprünglichen Gesicht haben, um gesichtsbiometrische Systeme auch ohne den Einsatz von Echtzeit-Deepfake-Videotechnologie zu überwinden.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Bei der Erstellung von gefälschten Dokumenten kommen Gesichtsgeneratoren und Gesichtsmorphing zum Einsatz. Zur Überwindung unternehmensseitiger Sicherheitsmassnahmen wird auf Echtzeit-Deepfake-Videotechnologie gesetzt. Beides erfordert derzeit einen hohen Ressourcenaufwand.

#### Adressatenkreis und Verbreitungsweise

Es gibt in diesem Szenario lediglich einen kleinen Adressatenkreis in Form der jeweils angegriffenen Unternehmen.

### A.6.2.2. Szenario 3: Rufschädigung eines Unternehmens mittels eines Deepfakes

#### Szenarioübersicht

Ein Rufschädigungsangriff kann grundsätzlich auf zwei Weisen durchgeführt werden. Bei der ersten Variante geraten das Unternehmen selbst, seine Produkte oder Führungspersonen ins Visier der Angreifer, indem mittels eines Deepfakes irreführende Informationen über das Unternehmen (z.B. fabrizierte Meldungen über einen Firmenkonkurs), seine Produkte (z.B. fabrizierte Meldungen

über gesundheitsschädigende Stoffe) oder Führungspersonen (z.B. fabrizierte Nachrichten über Korruption) verbreitet werden. Ein Topmanager einer Bank könnte dann etwa trotz einer angespannten Wirtschaftssituation über Sonderzulagen sprechen. Vertreterinnen oder Vertreter eines Pharmaunternehmens könnten über vermeintlich bekannte Nebenwirkungen von Medikamenten sprechen, die bewusst ignoriert würden.

Vergleichbare Fälle hat es bereits gegeben, etwa bei Amazon mit Fake-Profilen, die Deepfake-Fotos als Profilbild verwendeten. Indem in Anlehnung an eine gescheiterte PR-Kampagne Amazons die Arbeitspraktiken des Konzerns in übertriebener Weise gelobt wurden, sollten ebenjene Praktiken kritisiert werden (Hao 2021). Auch wenn die Kritik an den Arbeitsbedingungen bei Amazon angesichts regelmässiger Negativschlagzeilen (Toler 2019) vergleichsweise angemessen erscheint, sollte klar sein, dass derartige Deepfake-Profil-Angriffe gegen jedes beliebige Unternehmen möglich sind und auch in weniger gerechtfertigten Fällen Anwendung finden könnten.

Möglich wären auch kombinierte Angriffe, wenn beispielsweise zunächst fabrizierte Meldungen über einen Konkurs verbreitet werden und anschliessend dem CEO in einem fabrizierten Video die Bestätigung der Konkursmeldung in den Mund gelegt wird. Auch sind ebenenübergreifende Angriffe denkbar, indem z.B. mittels Cybererpressung erbeutete Informationen für eine grössere Verleumdungskampagne eingesetzt werden.

Zur Steigerung des Schadenspotenzials können synthetische Social Botnets verwendet werden, um eine echte oder fabrizierte Meldung an ein grösseres Publikum auszusenden. Schon heute existieren fabrizierte Social-Media-Accounts, die für koordinierte Kampagnen für oder gegen bestimmte Produkte eingesetzt werden. Zumeist operieren sie unauffällig und losgelöst voneinander und entfalten ihre Wirkung erst dann, wenn eine koordinierte Kampagne gestartet wird.

Die Herbeiführung von Panikkäufen (oder etwa eines Banken-Runs) kann sowohl mittels der Erstellung fabrizierter Meldungen (über ein Unternehmen, dessen Produkte oder Führungspersonen) erfolgen als auch durch die Verbreitung echter oder fabrizierter Meldungen mittels Social Botnets. Zudem ist auch bei dieser Variante die Kombination beider Angriffsformen denkbar, also qualitativ hochwertige fabrizierte Meldungen, die von einem Botnet verbreitet und potenziert werden.

### Angriffstyp und erforderlicher Ressourcenaufwand

Grundlage für dieses Szenario ist die Generierung rufschädigender Deepfakes, die sich direkt gegen ein Unternehmen, dessen Produkte oder Führungspersonen richten. Dazu eignen sich insbesondere gefälschte rufschädigende Aussagen und Handlungen. Etwa ein Unternehmensvideo, das derart manipuliert wurde, sodass die im Video abgebildeten Mitarbeitenden rufschädigende Aussagen treffen. Führungspersonen könnten in Deepfake-Videos beliebige diskreditierende Aussagen in den Mund gelegt werden. Das Sammeln von Medienmaterial über medienaffine Führungspersönlichkeiten wäre unkompliziert, im Falle medienscheuer Führungspersonen würde jedoch ein höherer Ressourceneinsatz erforderlich sein. Annähernd lippensynchrone Deepfake-Videos, in denen einer Zielperson gefälschte Aussagen in den Mund gelegt werden, erfordern Aufwand und Know-how.

Mittels Deepfake-Bilder könnten z.B. auch einzelne Produkte eines Unternehmens in rufschädigender Weise dargestellt werden. In der Zukunft wird es voraussichtlich auch möglich werden, nicht nur menschliche Gesichter, Sprache und Körperbewegungen, sondern vollständige Situationen synthetisch in Videoform zu generieren. Das Einsatzpotenzial solcher vollständig synthetischen Inhalte ist extrem vielfältig und kaum überschaubar.

Allerdings können auch Deepfakes mit niedriger Qualität zur Rufschädigung führen, etwa wenn ein gekennzeichnetes satirisches Deepfake beabsichtigter- oder unbeabsichtigterweise ein Unternehmen, eine Führungsperson oder ein Produkt diskreditiert.

### Adressatenkreis und Verbreitungsweise

Weil die Wirksamkeit eines Angriffs zu Rufschädigungszwecken grundsätzlich mit der Menge der Menschen korreliert, die den Inhalt betrachten und ihm Glauben schenken, handelt es sich um einen möglichst weiten Adressatenkreis. Um möglichst viele Menschen zu erreichen, würden etwaige Deepfakes also über alle einem Angreifer bzw. Verursacher zur Verfügung stehenden Kanäle verbreitet werden. Ressourcenstarke Angreifer könnten zur Steigerung der Reichweite auch auf synthetische Social Botnets zurückgreifen.

Je nach Kontext könnte es auch darum gehen, einen spezifischen Adressatenkreis zu erreichen, der beispielsweise als besonders anfällig für bestimmte Botschaften gilt. In diesem Fall könnte ein Deepfake zunächst über private oder halböffentliche Kommunikationskanäle (wie Chatgruppen) verbreitet werden. Dadurch würde zugleich die Wahrscheinlichkeit, dass der Deepfake vom betrof-

fenen Unternehmen bzw. der betroffenen Person und von Nachrichtenmedien wahrgenommen wird, reduziert, sodass die Einleitung von Gegenschritten verzögert wird. Spezifische Bevölkerungsgruppen könnten auch mittels Micro-Targeting und der Einbettung des Deepfakes in Werbeinhalte erreicht werden.

### A.6.2.3. Szenario 4: Initiierung von Finanztransaktionen

#### Szenarioübersicht

Wie mehrfach demonstriert, können Deepfakes für Social-Engineering-basierte Angriffe verwendet werden. Im bekanntesten Fall dieser Art wurde ein Deepfake-Audio 2019 erfolgreich dazu benutzt, einem Mitarbeiter eines britischen Energieunternehmens vorzugaukeln, dass der CEO des deutschen Mutterunternehmens am Telefon wäre. Auf Drängen des falschen Chefs hin, wonach eine zeitkritische Überweisung an einen ungarischen Zulieferer erforderlich sei, initiierte der Mitarbeiter die gewünschte Überweisung in Höhe von 220 000 Euro. Dabei wurde auf eine mittels Deepfake-Technologie erstellte Sprachaufnahme des CEO zurückgegriffen.

Mit zunehmender Verfügbarkeit von Echtzeit-Deepfakes wären vergleichbare Angriffe auch per Videoanruf denkbar. Mitarbeitende könnten dazu im ersten Schritt per Deepfake-Sprachanruf kontaktiert und im Falle von Skepsis aufseiten des Mitarbeitenden per Videoanruf kontaktiert werden.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Die Manipulation einer Zielperson zu bestimmten Handlungen (Social Engineering) basiert in diesem Szenario auf der Imitation von Stimme und ggf. Aussehen eines Vorgesetzten. Das Klonen einer Stimme ist heute schon relativ einfach möglich. Das Imitieren von Stimme und Gesicht in Echtzeit erfordert deutlich höhere Ressourcen.

#### Adressatenkreis und Verbreitungsweise

Weil ein solcher Angriff auf der Ausnutzung der Hilfsbereitschaft, des Vertrauens, der Angst oder des Respekts vor dem hierarchischen Gegenüber gründet, wäre der Adressatenkreis so eng wie möglich, um einer Entdeckung zu entgehen. Im Regelfall dürfte lediglich eine einzelne Zielperson über einen direkten Kanal (Anruf, Text- oder Sprachnachricht) kontaktiert werden. Hierarchische

Unternehmensstrukturen können in diesem Szenario zudem als verstärkender Faktor wirken, wenn Mitarbeitende ihre Zweifel an der Authentizität unausgesprochen lassen oder Zweifel erst gar nicht zulassen.

### A.6.2.4. Szenario 5: Deepfake-basierter Ransomware-Angriff zur Erpressung von Geld

#### Szenarioübersicht

Ransomware-Angriffe sind inzwischen die häufigste Form von Cyberangriffen. Dabei werden die Computersysteme eines Ziels zunächst mittels Schadsoftware infiltriert und lahmgelegt. Anschliessend wird das Unternehmen vor die Wahl gestellt, eine Lösegeldsumme zur Freischaltung der lahmgelegten Systeme zu überweisen oder die Veröffentlichung der erbeuteten Daten in Kauf zu nehmen. Teilweise werden erbeutete Daten trotz einer Lösegeldzahlung veröffentlicht, um den Gewinn zu maximieren.

Der häufigste Weg zum Einschleusen von Schadsoftware ist die Ausnutzung menschlicher Fehlbarkeit, indem an die Mitarbeitenden eines Zielunternehmens massenhaft E-Mails versendet werden. Die Schadsoftware wird freigesetzt, sobald eine Mitarbeitende z.B. einen Anhang öffnet (etwa eine Word-Datei), der einen schädigenden Makro enthält. Angesichts der aus Ransomware-Attacken resultierenden enormen Gefahren werden seitens der Unternehmen besondere Anstrengungen unternommen, die Mitarbeitenden darin zu schulen, derartige E-Mails zu erkennen und sie im Verdachtsfall an zuständige IT-Stellen zu melden. Mittels Deepfakes könnten derartige Gegenstrategien unterminiert werden. Hier sind verschiedene Angriffswege denkbar.

Kombiniert mit einem Deepfake-Video- oder Sprachanruf eines imitierten Vorgesetzten könnten Mitarbeitende dazu aufgefordert werden, einen in einer vorherigen Mail versendeten Schadinhalt zu öffnen und auszuführen.

Denkbar wäre auch der Versand einer zweiten gefälschten Mail im Namen des IT-Services, der die Unbedenklichkeit der zuvor versendeten Mail mit Schadinhalt bescheinigt bzw. zur Ausführung des Anhangs auffordert. Um etablierte, sichere Kommunikationskanäle zu umgehen, könnten ressourcenstarke Angreifer einen Videokonferenzlink in die gefälschte Mail des IT-Services hinzugeben, sodass Mitarbeitende im Falle von Rückfragen beim Angreifer landen, der die IT-Mitarbeitenden mittels Echtzeit-Deepfake imitiert.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Dieses Szenario basiert auf einer Mischung aus einem Social-Engineering-Angriff und gefälschten Aussagen und Handlungen, d.h., der erforderliche Ressourcenaufwand ist eher grösser.

#### Adressatenkreis und Verbreitungsweise

Der Adressatenkreis hängt von den Eigenschaften des Zielunternehmens ab. Häufig werden bei Ransomware-Angriffen E-Mails an möglichst viele Mitarbeitende versendet, um die Erfolgswahrscheinlichkeit zu steigern. Weil die oben beschriebene Kombination eines Ransomware-Angriffs mit Deepfakes allerdings das Vorhalten von zusätzlichen Ressourcen (Personen, die im Deepfake-Videoanruf Vorgesetzte oder IT-Mitarbeitende imitieren) erfordert, könnte der Adressatenkreis verringert werden, um beispielsweise die Gefahr der Aufdeckung durch Kontaktierung der echten IT-Stelle zu entgehen.

### A.6.2.5. Szenario 6: Abgreifen von vertraulichen Informationen (Wirtschafts- und Industriespionage)

#### Szenarioübersicht

Angesichts der Stärke und Innovationsfähigkeit der Schweizer Wirtschaft sind Schweizer Unternehmen Ziel von Wirtschafts- und Industriespionage. Die meisten Angriffsversuche basieren auf E-Mails mit kritischem Inhalt, also insbesondere Ransomware.

So kann das Abgreifen vertraulicher Informationen zum einen auf Social-Engineering-basierten Angriffen aufbauen: Die im vorherigen Szenario zur Erpressung von Geld mittels Ransomware-Attacke beschriebene Vorgehensweise kann auch zur Erbeutung von vertraulichen Informationen wie Betriebsgeheimnissen genutzt werden. Kombiniert mit einem Deepfake-Video- oder Sprachanruf eines imitierten Vorgesetzten könnten Mitarbeitende dazu aufgefordert werden, einen in einer vorherigen Mail versendeten Schadinhalt zu öffnen und auszuführen, sodass IT-Systeme infiltriert werden.

Zum anderen können mittels Deepfake-Technologie auf Stimm- und Gesichtserkennung basierende Sicherheitssysteme getäuscht werden, um mittels Identitätsdiebstahls Zugang zu IT-Systemen und damit zu vertraulichen Informationen zu erhalten.

### Angriffstyp und erforderlicher Ressourcenaufwand

Bei diesem Szenario ist Identitätsdiebstahl die Grundlage für zwei Angriffstypen. Indem unter hohem Ressourceneinsatz das Aussehen und die Stimme z.B. eines Mitarbeitenden imitiert werden, könnten Angreifer erstens Zugang zu stimm- und gesichtsbiometrisch geschützten Bereichen erhalten. Dies könnte beispielsweise eine physische Zugangskontrolle zu einem Gebäude oder Raum sein und/oder der Zugang zu einem digitalen Endgerät wie einem Computer oder Smartphone.

Indem Aussehen und die Stimme einer Person imitiert werden, könnten, zweitens, mittels Social Engineering Zielpersonen getäuscht und zur Durchführung bestimmter Handlungen manipuliert werden, beispielsweise indem ein imitierter Vorgesetzter die Herausgabe vertraulicher Informationen verlangt oder die Ausführung von E-Mail-Anhängen anordnet, sodass Schadsoftware installiert wird. Das Klonen der Stimme ist vergleichsweise einfach, die Imitation von Stimme und Gesicht in Echtzeit hingegen deutlich aufwendiger.

### Adressatenkreis und Verbreitungsweise

Auch in diesem Szenario hängt der Adressatenkreis von den Eigenschaften des Zielunternehmens ab. Bei Ransomware-Angriffen werden E-Mails häufig an möglichst viele Mitarbeitende versendet, um die Erfolgswahrscheinlichkeit zu steigern. Weil die oben beschriebene Kombination eines Ransomware-Angriffs mit Deepfakes allerdings das Vorhalten von zusätzlichen Ressourcen (Personen, die im Deepfake-Videoanruf Vorgesetzte oder IT-Mitarbeitende imitieren) erfordert, könnte der Adressatenkreis verringert werden, um beispielsweise die Gefahr der Aufdeckung durch Kontaktierung der echten IT-Stelle zu entgehen.

Im Falle der Überwindung von stimm- und gesichtsbiometrischen Sicherheitsmassnahmen ist das Angriffsziel kein Mensch, sondern in aller Regel ein Log-in-System.

## A.6.2.6. Szenario 7: Onlinewerbebetrug mittels synthetischer Profile

### Szenarioübersicht

Spezialisierte Unternehmen bieten seit einigen Jahren ihre Dienste zum Zwecke der Vergrösserung der Followerschaft oder der Nutzerinteraktion auf sozialen Netzwerken an. Dazu werden gefälschte Nutzerprofile auf soziale Medienplattformen erschaffen, die gegen eine entsprechende Gebühr automatisiert

den Käufern auf den Plattformen folgen und mit ihren Posts interagieren oder auch bestimmte Webseiten ansteuern, sodass diese als erfolgreicher wahrgenommen werden, als sie es in Wirklichkeit sind. Insbesondere würden derartige Dienste seitens Influencern wahrgenommen, um Werbebetrug zu betreiben und ihre Werbeeinnahmen zu steigern. Bei rund einem Viertel aller Instagram-Influencer seien ca. 15 % der Followerschaft über derartige Anbieter erkauft (Cavazos 2019). Unternehmen, die auf Influencer setzen, um ihre Produkte zu vermarkten, können daraus insofern einen unmittelbaren wirtschaftlichen Schaden erleiden, dass sie angesichts künstlich erhöhter Follower- und Interaktionszahlen mehr Geld für Werbekampagnen ausgeben, als angesichts der realen Zahlen angemessen wäre. Schätzungen zufolge entstünden werbetreibenden Unternehmen auf diese Weise weltweit Schäden in Höhe mehrerer Milliarden US-Dollar (University of Baltimore 2020).

Die auf Fake-Follower- und -Interaktionen spezialisierten Unternehmen setzen auf immer bessere Mittel, um ihre Fake-Nutzerprofile vor den Plattformbetreibern und sonstigen Detektoren zu verschleiern. Deepfakes können in diesem Zusammenhang ein Mittel sein, um authentisch wirkende künstliche Nutzerprofile zu erschaffen, die nicht oder nur schwer von Plattformbetreibern entdeckt werden können.

**Angriffstyp und erforderlicher Ressourcenaufwand**

So könnten Deepfake-Gesichtsgeneratoren dazu eingesetzt werden, «neue» Gesichter von Menschen zu generieren, die in der Realität nicht existieren. Auf diese Weise können beispielsweise Detektoren, die auf die Erkennung von gefälschten Profilen mittels Profilfotoabgleichs ausgelegt sind, effektiv umgangen werden. Dies ist mit sehr geringem Ressourcenaufwand möglich. Und auch KI-generierte Texte bzw. die fortschrittliche Chat-Funktionalität von modernen Chatbots (wie z.B. ChatGPT) können dazu verwendet werden, authentisch wirkende Interaktionen der Fake-Profile vorzutäuschen, sodass eine Unterscheidung zwischen echten Profilen und Fake-Profilen zunehmend schwieriger wird. Ein derartiger Missbrauch von Textgeneratoren ist zwar aufwendig, angesichts des bereits heute bestehenden Schwarzmarkts, auf dem Dienstleistungen zur Durchführung von Werbebetrug angeboten werden (Cavazos 2019; Aznar 2021), ist die Entwicklung entsprechender neuer Produkte seitens professioneller Akteure allerdings nicht unwahrscheinlich.

**Adressatenkreis und Verbreitungsweise**

Dadurch, dass gefälschte Profile möglichst authentisch mit anderen Profilen interagieren müssen, um nicht als Bot wahrgenommen zu werden, kann nicht von einem Adressatenkreis wie in anderen Szenarien gesprochen werden. Die Verbreitung etwaiger synthetischer Inhalte erfolgt mittels der sozialen Medien, auf denen die Bots aktiv sind.

### A.6.3. Marktebene

#### A.6.3.1. Szenario 8: Marktmanipulation

**Szenarioübersicht**

Auf systemischer Ebene werden mehrere Einsatzmöglichkeiten von Deepfakes diskutiert, darunter insb. die Manipulation des (Aktien-)Marktes sowie die Auslösung und Verschärfung eines kurzfristigen Wirtschaftseinbruchs.

Hierbei ist insbesondere an fabrizierte Nachrichtenmeldungen zu denken, bspw. über einen terroristischen Angriff auf Regierungsangehörige oder über eine vermeintliche staatliche Investigation eines Wirtschaftszweigs, die einen kurzfristigen Wirtschaftscrash auslösen können. Einen einschlägigen Präzedenzfall bildet die Übernahme des Twitter-Accounts der Associated Press durch die sog. Syrian Electronic Army im April 2013. Als der gekaperte Associated Press-Account einen Tweet absetzte, in dem Bombenexplosionen im Weissen Haus und die Verletzung des damaligen US-Präsidenten Barack Obama behauptet wurden, verlor der S&P 400-Aktienindex innerhalb von drei Minuten an Wert in Höhe von 136 Milliarden US-Dollar. Zwar wurde die Meldung schnell als Falschmeldung entlarvt und der eingestürzte Markt erholte sich innerhalb von nur drei weiteren Minuten, doch bestand genug Zeit für Angreifer, um Profite mit den gefallenen Kursen zu erwirtschaften (Foster 2013). Ein solcher Angriff könnte auch mittels eines Deepfake-Videos ausgeführt werden, um deren Glaubwürdigkeit und damit Wirksamkeit zu steigern.

Besonders gefährlich könnte ein solcher Deepfake sein, wenn er während einer aufgeladenen gesellschaftlichen Situation gestreut wird. Eine solche Situation hat es beispielsweise während des Liquiditätsengpasses der Credit Suisse Anfang 2023 gegeben (Bartz 2022). Ein Deepfake, der in einer vergleichbaren Situation veröffentlicht wird und den Beinahekollaps weiterer wichtiger Banken zum Inhalt hat oder der politische Verantwortliche vermeintlich dabei zeigt, wie

sie einen kurzfristigen politischen Wirtschaftseingriff ankündigen, könnte auch in der Schweiz einen marktmanipulierenden Effekt haben.

Weniger wahrscheinlich ist, dass mittels eines Deepfakes ein grossangelegter Angriff auf das gesamte Schweizer Wirtschaftssystem erfolgt. Als anfällig für derartige Angriffe gelten insbesondere Staaten mit einer instabilen Wirtschaft und einem geringen Vertrauen der Bevölkerung in die Wirtschaft (Bateman 2020: 24). Angesichts der stabilen wirtschaftlichen Lage der Schweiz und des im internationalen Vergleich relativ hohen Vertrauens der Schweizer Bevölkerung in das Wirtschaftssystem (Enste/Suling 2020: 18) kann also grundsätzlich davon ausgegangen werden, dass Marktmanipulation vor allem als Ad-hoc-Angriff eingesetzt werden könnte.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Der wahrscheinlichste Angriffstyp ist die Fälschung von Aussagen oder Handlungen mit dem Ziel der Irreführung der Gesellschaft und insb. von Marktteilnehmern. Unter hohem Ressourcenaufwand könnten synthetische Social Botnets zudem eine wichtige Rolle bei der schnellen und effektiven Verbreitung des Deepfake-Inhalts spielen.

#### Adressatenkreis und Verbreitungsweise

Zum Zwecke der Marktmanipulation eingesetzte Deepfake-Inhalte werden voraussichtlich einen möglichst weiten Adressatenkreis haben. Weil davon auszugehen ist, dass eine Gegenreaktion bzw. Klarstellung auf einen Deepfake schnell erfolgen würde, würde ein solcher Deepfake – ähnlich wie im oben geschilderten Fall der Vereinigten Staaten – möglichst schnell an möglichst viele Menschen ausgesendet werden. Das Kapern von Nachrichten- und Medienorganisationen erscheint kurzfristig auch als eine besonders gefährliche Angriffsform, weil so das Vertrauen der Bevölkerung und Marktteilnehmer in die Medien missbraucht wird, um einen schnellen und starken Effekt zu erzielen. Alternativ könnten synthetische Social Bots entsprechende Deepfake-Inhalte oder «Mundpropaganda» streuen.

### A.6.3.2. Szenario 9: Digitales Astroturfing: Beeinflussung von demokratischen Entscheidungsprozessen zulasten der Wirtschaft

#### Szenarioübersicht

Schliesslich kann **digitales Astroturfing** zur Beeinflussung politischer Prozesse zulasten von bestimmten Unternehmen und Wirtschaftssektoren verwendet werden. Dabei werden von Bots verbreitete KI-generierte Texte zur Überflutung einer öffentlichen Konsultation mittels gefälschter Positionierungen verwendet, um den Gesetzgebungsprozess in Richtung der gewünschten Position zu manipulieren (Bateman 2020: 25 ff.). Auch wenn erste empirische Untersuchungen des Phänomens in den Vereinigten Staaten auf eine eher vernachlässigbare Wirkung hindeuten, kann digitales Astroturfing politische Entscheidungsprozesse beeinflussen (Balla u.a. 2022; Handan-Nader 2023).

Insbesondere dann, wenn politische Entscheidungstragende und Verwaltungsmitarbeitende in erster Linie auf die schiere Anzahl der unterstützenden oder ablehnenden Eingaben blicken. Aber auch dann, wenn bei politischen Entscheidungsprozessen, etwa weil es sich um ein öffentlich umstrittenes Thema handelt, weniger die Beachtung von Expertenwissen im Vordergrund steht, als die Gewinnung einer möglichst breiten gesellschaftlichen Unterstützung durch öffentliche Konsultationsprozesse (Handan-Nader 2023: 112). Zudem könne der geringe Einfluss auch auf die niedrige inhaltliche Qualität der gefälschten Positionierungen zurückzuführen sein (ebd.: 113). Die zu erwartende Zunahme der inhaltlichen Qualität synthetischer Textinhalte angesichts hochwertiger Chatbots könnte es also in Zukunft ermöglichen, umfangreiche und informative Fake-Positionierungen zu erstellen und in öffentliche Konsultationsprozesse einzuspeisen. Daher ist eine Steigerung des Gefahrenpotenzials in der Zukunft denkbar.

#### Angriffstyp und erforderlicher Ressourcenaufwand

Digitales Astroturfing basiert auf der Generierung synthetischer Inhalte mittels KI-generierter Texte und der Verbreitung der Inhalte entweder durch deren Eingabe direkt in einen öffentlichen Konsultationsprozess oder der Generierung öffentlicher Aufmerksamkeit durch massenhafte Streuung entsprechender Inhalte über soziale Medien. Im Zusammenhang mit Letzterem könnten ausserdem synthetische Social Bots eine wichtige Rolle spielen. Jeder der Schritte erfordert einen hohen Ressourcenaufwand.

### Adressatenkreis und Verbreitungsweise

Wenn es das Ziel wäre, politische Entscheidungstragende in frühen Gesetzesphasen, etwa während der Formulierung eines Vorschlags zu beeinflussen, würde eine Einflussnahme auf den engen Adressatenkreis der entsprechenden Entscheidungstragenden fokussieren. Typische Kommunikationskanäle wären entweder Einreichungen während einer formellen Konsultationen oder direkte E-Mail-Kommunikation.

Wenn die Einflussnahme in einer späteren Phase des Gesetzgebungsprozesses erfolgt, könnte als Hebel zur Beeinflussung der Entscheidungstragenden die breite Öffentlichkeit genutzt werden. In diesem Fall würden soziale Medienplattformen, Chatgruppen usw. als Mittel verwendet werden, um eine möglichst breite öffentliche Aufmerksamkeit auf ein Thema zu lenken, indem Emotionen zu dem Thema erzeugt werden.

Ein weiter Adressatenkreis wäre auch dann relevant, wenn wichtige Entscheidungspersonen mittels diskreditierender Deepfakes aus dem Amt gedrängt werden sollen. Im Falle der Beeinflussung der individuellen Entscheidung von Amtsträgern oder Parlamentariern würden diese Personen mittels erpresserischer und einschüchternder Deepfakes direkt kontaktiert werden.

# 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**, Redaktor, Schweizer Radio und Fernsehen SRF, 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

www.ta-swiss.ch

mitglied der
akademien der
wissenschaften schweiz

*Die Imitation von Stimmen und Gesichtern mittels neuer KI-Technologien wird zunehmend einfacher – und die Resultate sind schon heute kaum mehr von echten Stimmen und Gesichtern zu unterscheiden. Diese Technologien bieten viele kreative und innovative Nutzungspotenziale, sowohl für Privatnutzende als auch für verschiedene Wirtschaftsbereiche. Andererseits sind Deepfake-Videos und Deepfake-Audios auch Quell zahlreicher Risiken: Sie können z.B. politische Desinformation vereinfachen, Mobbing Vorschub leisten und kriminelle Aktivitäten begünstigen.*

*Angesichts der fortwährenden Debatten über die Chancen und Risiken von Deepfakes wird in der Studie der Umgang mit einer veränderten Realität diskutiert, in der täuschend echte Fälschungen mit Originalmedien um die Aufmerksamkeit der Zuschauer konkurrieren. Was ist technisch möglich? Welche Grenzen setzt das Recht? Was weiss und denkt die Bevölkerung über Deepfakes? Wie geht der Journalismus mit dem Phänomen um? Welche Rolle könnten Deepfakes in Politik und Wirtschaft spielen? Die Studie bietet zu diesen und weiteren Fragen fundierte Orientierung und zeigt darauf basierende Handlungsmöglichkeiten auf.*

v/d/f

**TA-SWISS 81/2024**

**ISBN 978-3-7281-4185-9 (Printversion)**

**ISBN 978-3-7281-4186-6 (E-Book)**

**DOI-Nr. 10.3218/4186-6**