Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags

Murat Karaboga

1 Einführung und Problemstellung

Spätestens mit dem aufsehenerregenden Obama-Deepfake aus dem Jahr 2018 wurden die technologischen Möglichkeiten von Deepfakes einem größeren Zuschauer:innenkreis bekannt (BuzzFeed, 2018). Seither sind zwar glücklicherweise größere Deepfake-induzierte Skandale ausgeblieben.[1] Politiker:innen und Forschende weltweit arbeiten jedoch trotzdem mit Hochdruck an Regulierungsvorschlägen zur Einhegung von Deepfakes, um die befürchteten Effekte von mit Schädigungsabsicht erstellten und verbreiteten Deepfakes zu vermeiden.

Dabei sind Manipulationen von Bild- und Tonmaterial freilich nichts Neues. Seit es menschengemachte Aufzeichnungen gibt, existieren auch Fälschungen (Steinebach et al., 2020, S. 21). Frühere Manipulationsmethoden stützten sich allerdings vor allem auf Erzählungen und Geschriebenes und derartige Fälschungen von Bild- und Tonmaterial konnten auf relativ einfache Weise

[1] An dieser Stelle soll nicht unerwähnt bleiben, dass Deepfakes und insb. Deepfake-Pornographie dennoch bereits sehr reale Konsequenzen für die Opfer haben, wie z. B. der Fall Ayyub aus dem Jahr 2018 demonstriert (Ayyub, 2018).

M. Karaboga (✉)
Fraunhofer-Institut für System- und Innovationsforschung ISI, Competence Center Neue Technologien, Karlsruhe, Deutschland
E-Mail: Murat.Karaboga@isi.fraunhofer.de

S. Jaki und S. Steiger (Hrsg.), *Digitale Hate Speech*,
https://doi.org/10.1007/978-3-662-65964-9_10

wirksam erkannt werden. Erstmals ist es mittels KI-Einsatz inzwischen auch möglich, qualitativ hochwertige synthetische Medieninhalte zu erstellen oder bestehende Inhalte so zu manipulieren, dass sie sich zunehmend schwieriger von realen Inhalten unterscheiden lassen. Weil die menschliche Psyche dem, was die eigenen Augen sehen, stärkeren Glauben zu schenken scheint, wird Deepfakes ein besonders großes Manipulationspotential beigemessen (Minsky, 2021). Diese Herausforderung wird dadurch verstärkt, dass technologische Fortschritte zum einen die Qualität von Deepfake-Inhalten in absehbarer Zeit dahingehend steigern könnten, dass selbst Expert:innen nicht mehr in der Lage wären, eine Unterscheidung von realen Inhalten vorzunehmen (Jiang et al., 2021). Zum anderen führen ebenjene Fortschritte dazu, dass die Herstellung von täuschend echten Deepfakes immer einfacher wird. Deshalb wird davon ausgegangen, dass künftig auch technisch nicht- oder nur wenig versierte Menschen entsprechende Inhalte erstellen können werden (van Huijstee et al., 2021, S. 7 ff.).

Den mit Deepfakes assoziierten Risiken wird auch deshalb größere Aufmerksamkeit beigemessen, da das Deepfakes-Phänomen in eine *gefährliche Zeit* fällt, die dadurch charakterisiert ist, dass die Verbreitung medialer Inhalte und insbesondere von Nachrichteninhalten nicht mehr allein in der Hand von Akteuren liegt, die vertrauenswürdig sind bzw. deren Handlungen quantitativ überschaubar sind (Chesney & Citron, 2019, S. 9 f.). Durch den sogenannten *digitalen Strukturwandel* sind neben Massenmedien mit Onlineplattformen neue Vermittler öffentlicher Kommunikation getreten (Eisenegger, 2021). Plattformen wie soziale Netzwerke, Suchmaschinen, Video-Sharing-Dienste oder Nachrichtenaggregatoren spielen auch für die Nutzung herkömmlicher Medien eine zentrale Rolle. Von klassischen Blogs, über soziale Online Netzwerk-Plattformen bis hin zu Instant Messaging Apps bietet sich für all jene, die an der Verbreitung synthetischer oder manipulierter Inhalte interessiert sind, eine nie da gewesene Vielzahl an Verbreitungsmöglichkeiten. Wie schon aus der Diskussion rund um Desinformation und Social Media bekannt, treibt diese Entwicklung nicht nur die Quantität der Verbreitung entsprechender Inhalte voran. Im Umkehrschluss bedeutet die zunehmende Nachrichtennutzung auf sozialen Netzwerken, insbesondere bei jüngeren Altersgruppen, dass sowohl korrekte Informationen als auch Richtigstellungen über Desinformation ihre Adressaten häufig nicht erreichen. Hinzu kommt, dass psychologische Faktoren dazu führen, dass manipulierte Inhalte, gerade dann, wenn sie besonders sensationelle Informationen zu liefern vorgeben und/oder bestehende Einstellungen affirmieren (der sog. *Confirmation Bias*), viel stärker wirken, als Klarstellungen es vermögen. Dadurch hinterlassen entsprechende Inhalte selbst dann noch einen signifikanten

Effekt bei den Betroffenen, wenn die konsumierten Inhalte später richtiggestellt werden (der sog. *Falschinformationseffekt*) (Högden et al., 2020).

Angesichts dieser Problemlage werden weltweit Wege diskutiert, wie den möglichen Risiken von Deepfakes begegnet werden sollte. Unter den Vorschlägen finden sich solche zur Technologiegestaltung, zur Entwicklung von Detektionsmethoden, zum Ausbau der gesellschaftlichen Resilienz bis hin zu verschiedenen Möglichkeiten des Verbots oder der anderweitigen Begrenzung der Verbreitung manipulierter oder synthetischer Inhalte (Trend Micro Research, UNICRI und EC3, 2020, S. 61 ff.). Mit den gegenwärtig auf EU-Ebene verhandelten Vorschlägen zu neuen Regelungen für Online-Plattformen (DSA) und KI-Regulierung (AI Act) existieren zwei Vorhaben, die Deepfakes unmittelbar betreffen.

Der vorliegende Beitrag gibt einen Überblick über die Regulierung von Deepfakes auf EU-Ebene und diskutiert, inwiefern die Kommissionsentwürfe zum DSA und KI-Verordnungsvorschlag (AI Act) die mit Deepfakes assoziierten Herausforderungen adressieren. Der Aufsatz beginnt mit einem Kapitel zur begrifflichen Einordnung des Phänomens Deepfakes innerhalb des Diskurses rund um Desinformation, Fehlinformation, schädigende Informationen und Hate Speech. Das daran anschließende Kapitel gibt einen Überblick über die mit Deepfakes assoziierten Chancen und Risiken.

Daran schließt sich die Diskussion der Regulierungsbemühungen zu Deepfakes an. Nachdem ein Überblick über Regulierungsbestrebungen in den Vereinigten Staaten und der Volksrepublik China gegeben wurde, werden die einschlägigen Vorgaben und Regulierungsbestrebungen auf EU-Ebene vorgestellt. Der Fokus hierbei liegt auf den EU-Vorhaben zum Digital Services Act und der KI-Regulierung, die sowohl im Hinblick auf Berührungspunkte mit Deepfakes als auch im Hinblick auf mögliche Probleme diskutiert werden, sodass Handlungspotentiale deutlich werden.

2 Definitionen

Im Folgenden werden zentrale Begrifflichkeiten des Deepfakes-Diskurses definiert und zu einander in Beziehung gesetzt.

Deepfakes: Bei der Definition des Begriffs Deepfakes folgt der Beitrag dem Vorschlag einer im Auftrag des Europäischen Parlaments erarbeiteten Studie. Deepfakes werden darin als ***„manipulierte oder synthetische Audio- oder Video-Inhalte verstanden, die authentisch erscheinen und in denen (eine) Person(en)***

etwas zu sagen oder zu tun scheint (scheinen), was sie nie gesagt oder getan hat (haben), und die mithilfe von KI-Techniken, einschließlich maschinellem Lernen und Deep Learning, hergestellt wurden“ (van Huijstee et al., 2021, S. 2 – eigene Übersetzung).

Obwohl Videomanipulationen schon viel länger existieren als das Phänomen Deepfakes, rückt mit dem Begriff die neuartige technische Produktionsdimension in den Vordergrund, die es erstmals möglich macht, Videos von derart hoher und zunehmender Qualität – immer einfacher auch seitens Laien – zu produzieren, dass sie von echten Inhalten kaum mehr unterscheidbar sind. Die Definition lässt zudem offen, ob Deepfakes auch für nicht-schädliche (z. B. künstlerische) Zwecke verwendet werden können. ***Cheapfakes*** bezeichnen Deepfakes mit geringer Qualität, wie zum Beispiel das Nancy Pelosi-Video, das lediglich eine verlangsamte Version des Originalvideos darstellte, bei dem keine sonstigen, fortschrittlichen Manipulationstechniken Anwendung fanden (Paris & Donovan, 2019). Bislang knüpft die Definition von Deepfakes an die manipulierte Abbildung von Personen an. Es ist denkbar, dass künftige Deepfakes auch Objekte oder die Umwelt, beispielsweise fiktive Naturkatastrophen, abbilden, sodass eine Erweiterung der Definition auf den Einschluss nicht-personenbezogener Situationen geboten sein könnte (Centre for Data Ethics and Innovation, 2019).

Deepfakes können für beliebige Zwecke verwendet werden und stellen somit ein Querschnittsphänomen dar. In der einschlägigen Literatur werden Deepfakes – vor allem aufgrund der politischen Bedeutungszuschreibung – primär im Bereich von Desinformation und Fehlinformation verortet. Allerdings können Deepfakes ebenso zu Zwecken der schädigenden Information und Hate Speech eingesetzt werden (vgl. Ajder et al., 2019; Bateman, 2020; Chesney & Citron, 2019). Daher werden diese Begriffe im Folgenden definiert.

Desinformation: Desinformation wird in Anlehnung an den Formulierungsvorschlag der hochrangigen Expertengruppe für Fake News[2] und Desinformation der Europäischen Union (EU) definiert. Demnach beinhaltet Desinformation „nachweislich falsche oder irreführende Informationen, die mit dem Ziel des wirtschaftlichen Gewinns oder der vorsätzlichen Täuschung der Öffentlichkeit konzipiert, vorgelegt und verbreitet werden und öffentlichen Schaden anrichten

[2] Der noch vor einigen Jahren äußerst populäre Begriff „Fake News“ wird in der wissenschaftlichen und politischen Debatte inzwischen vermieden, da er inhaltlich unscharf und zugleich politisch aufgeladen ist (Wardle & Derakhshan, 2017, S. 5).

können. […] Irrtümer bei der Berichterstattung, Satire und Parodien oder eindeutig gekennzeichnete parteiliche Nachrichten oder Kommentare sind keine Desinformation.“ (Europäische Kommission, 2018, S. 4).

Fehlinformation (Misinformation): Fehlinformationen werden verstanden als Informationen, „die faktisch nicht der Wahrheit entsprechen, jedoch vom Sender, der die Information verbreitet, für wahr gehalten werden“ (Johann & Wagner, 2020, S. 102).

Schädigende Information (Malinformation): Schädigende Informationen sind von Desinformation und Misinformation insofern zu unterscheiden, als sie der Wahrheit entsprechen, ihre Verbreitung allerdings der gezielten Schädigung bzw. Benachteiligung der Zielpersonen dient. Beispiele sind Leaks und Doxing (Wardle & Derakhshan, 2017, S. 5).

Hate Speech: Unter Hate Speech werden im Folgenden Nachrichten verstanden, die sich gegen ein Individuum oder eine Gruppe richten, indem Identitätsmerkmale (Geschlecht, Ethnie, Hautfarbe, philosophische Überzeugungen oder Religion, Nationalität) des Individuums bzw. der Gruppe als negativ und unerwünscht dargestellt und die Abwertung eines Individuums bzw. einer Gruppe beabsichtigt wird. Dadurch kann Hate Speech auch zu körperlicher Gewalt führen (Rudnicki & Steiger, 2020).

3 Mit Deepfakes assoziierte Risiken und Chancen

Deepfake-Technologien sind eine typische *dual use*-Technologie. Sie können sowohl für gesellschaftlich erwünschte als auch für gesellschaftlich unerwünschte Zwecke eingesetzt werden. Jegliche Regulierung ist daher gefordert, bei dem Versuch der Verhinderung von schädlichen Effekten erwünschte Potentiale von Deepfakes nicht zu ersticken.

3.1 Chancen von Deepfakes

Deepfake-Technologien ermöglichen eine Reihe von erwünschten Nutzungsmöglichkeiten. Zu den bekanntesten dieser Anwendungen zählen die Verwendung von Smartphone-Kamera-Apps und die Wiederbelebung verstorbener bzw. Verjüngung gealterter Schauspieler:innen. Das Austauschen von Gesichtern (sog.

Face-Swap-Videos) mittels Smartphone-Apps zu Unterhaltungszwecken ist in den vergangenen Jahren, insbesondere unter Jugendlichen, zu einer gängigen Kulturpraxis avanciert. Anwendungen wie Snapchat und Reface zählen zu den weltweit am häufigsten heruntergeladenen Apps (Curry, 2021). Entsprechend ist es möglich, dass Menschen sich selbst in kurze Schnipsel aus Hollywood-Filmen *hineindeepfaken* (Kietzmann et al., 2020). Disney verwendete Deepfake-Technologie zunächst zur Darstellung einer Rolle, deren Besetzung verstorben war, und später auch zur Inszenierung des jungen Luke Skywalkers auf der Kino- bzw. Streaming-Leinwand.

Generell eröffnen Deepfake-Technologien zahlreiche Nutzungsmöglichkeiten für Audio-, Foto- und Videoproduzenten bzw. -bearbeiterinnen: die einfache und kostengünstige Korrektur falsch gesprochener Zeilen durch Deepfake-Audio; die vereinfachte Erstellung von 3D-Modellen in Computerspielen; der Einsatz im Kulturbereich, bspw. zur Nachstellung historischer Ereignisse inklusive der Darstellung verstorbener historischer Persönlichkeiten uvm. (van Huijstee et al., 2021, S. 26–29).

Chancen von Deepfakes werden auch zur Verbesserung der Mensch-Maschine-Interaktion, für Videokonferenzen, bei medizinischen Anwendungen sowie zu Satirezwecken gesehen. Insbesondere satirische Verwendungen sind im Internet vielfach anzutreffen und können zu generellen satirisch-humoristischen Zwecken eingesetzt werden. Ein Beispiel hierfür ist etwa die Parodie der Weihnachtsansprache von Königin Elizabeth II. im Jahr 2020 (Channel 4, 2020).

3.2 Risiken von Deepfakes

Ein großer Teil der Publikationen zu Deepfakes verstehen diese als eine Risiko-Technologie, deren Wirkungen sich, sofern nicht rechtzeitig und mit angemessenen Mitteln gegengesteuert wird, erst in einigen Jahren entfalten werden (Bateman, 2020; Chesney & Citron, 2019; Collins, 2019; Dobber et al., 2021; Trend Micro Research, UNICRI und EC3, 2020, S. 60 f.). Diskursiv ist das Thema Deepfakes an der Schnittstelle zwischen den Debatten rund um Desinformation einerseits und Hate Speech andererseits zu verorten. Mit der Dimension der Desinformation werden insbesondere gesellschaftliche Effekte assoziiert, die noch nicht eingetreten sind, aber als Zukunftsszenario befürchtet werden. Dazu zählen beispielsweise die Manipulation von Wahlen, die generelle Erosion gesellschaftlichen Vertrauens oder auch ein sinkendes Vertrauen in das Justizsystem. Gleichzeitig demonstriert der Status Quo der Deepfake-Nutzung, dass der (rache-)pornographische Einsatz von Deepfakes am häufigsten anzu-

treffen ist. So machten pornographische Inhalte (Deepfake-Pornographie oder Deepnudes), die fast ausschließlich Frauen abbildeten, im Jahr 2019 Schätzungen zufolge über 90 % aller schädigenden Deepfakes aus. Die Mehrzahl dieser Inhalte (über 90 %) bezieht sich zwar wiederum auf die Verbreitung pornographischer Inhalte auf spezifisch zur Verbreitung dieser Pornographie-Art ausgelegten Webseiten, sodass meist ein Publikum mit einem spezifischen Interesse an dieser Art von Inhalten auf diese stößt (Ajder et al., 2019, S. 1).[3] Allerdings zirkulieren sie in zunehmenden Maße auch über die Grenzen dieser Webseiten hinaus und werden zu Zwecken der Verleumdung, Bedrohung oder Erpressung von Frauen verwendet (Compton, 2021). Dies verdeutlicht, dass trotz des diskursiven Fokus' auf die Dimension der Desinformation in der einschlägigen Literatur gegenwärtig vor allem die Hate-Speech-Dimension von Deepfakes die größere Rolle zu spielen scheint.

Ein beliebter Einsatz zur Typologisierung der Risiken von Deepfakes ist die Unterteilung in Schadensarten: *Finanzieller Schaden*, *Reputationsschaden* und *Manipulation der Entscheidungsfindung* (Collins, 2019). Risiken können aber auch nach den Betroffenen strukturiert werden. Bateman (2020) sowie Chesney und Citron (2019) schlagen eine Unterteilung in die Kategorien (a) Individuen, (b) Institutionen bzw. Organisationen und (c) gesellschaftliche Schäden vor. Die Schwierigkeit bei letzterem Ansatz ist, dass viele der möglichen Schäden kaskadierende Effekte entfalten können und über das Individuum hinaus auch zu organisationalen und gesellschaftlichen Problemen führen können. Beispielsweise betrifft die Diskreditierung von Journalist:innen mittels Deepfakes in erster Linie die betroffene(n) Person(en), doch können die Reaktionen auf das Deepfake zusätzlich auf die Medienanstalt zurückfallen, bei der die Person(en) beschäftigt ist/sind. Je nach gesellschaftlicher Ausgangssituation und sonstigen Kontextfaktoren wie der Häufigkeit und des Schweregrads des Deepfakes *(wird im Deepfake beispielsweise eine systematische Nachrichtenfälschung unterstellt?)* können sich schließlich kaskadierende Effekte auf gesellschaftlicher Ebene zeigen, die beispielsweise zu einer Beschädigung des Vertrauens in das Mediensystem führen. Dementsprechend folgt der Beitrag der Sortierung nach Schadensarten.

Psychologische Schäden können durch Deepfakes insbesondere bei den von einem Deepfake unmittelbar betroffenen Personen entstehen. Deepfakes

[3] In sog. Deepfake-Pornos werden die Gesichter beliebiger Menschen in existierende pornographische Videos (teils auch Fotos) übertragen. Bei Deepnudes wird die auf einer Fotografie getragene Kleidung unter Software-Einsatz automatisch vollständig entfernt.

können etwa zum Zwecke des *Mobbings,* der *Verleumdung* und *Einschüchterung* eingesetzt werden. Wie erwähnt, haben Deepfakes eine starke sexuelle und sexistische Komponente, die neben den bereits genannten Zwecken auch zu *Rufschädigung* und zur *Unterdrückung der Freiheit der Meinungsäußerung* führen kann. Des Weiteren lassen sich Deepfakes zu Erpressungszwecken einsetzen. Wenn pornographische Deepfakes für Erpressungen genutzt werden, ist von *Sextortion* die Rede (van Huijstee et al., 2021, S. 30).

Finanzielle Schäden können durch Deepfakes sowohl bei Individuen als auch bei Organisationen entstehen. So können *Erpressungen* neben dem oben erwähnten psychologischen Schaden auch finanziellen Schaden für Individuen und Organisationen bewirken. Darüber hinaus können Deepfakes zum Zwecke des *Identitätsdiebstahls* eingesetzt werden. Ziele dieser Art des Betrugs können insbesondere Organisationen bzw. Unternehmen sein (Bateman, 2020, S. 9–11), aber auch Einzelpersonen (etwa in einer neuen Form des Enkel-Tricks) (Sokolov et al., 2020, S. 513–16). Zukünftig ist auch denkbar, dass Deepfakes zum Zwecke der Marken- oder Rufschädigung eingesetzt werden, indem beispielsweise Geschäftsführende dabei abgebildet werden, wie sie falsche Aussagen über Geschäftszahlen und Konkurse treffen. Je nach Reaktionsgeschwindigkeit des Unternehmens und Tragweite des Deepfakes kann es auch zu Manipulationen des Aktienmarktes kommen (van Huijstee et al., 2021, S. 31).

Gesellschaftliche Schäden durch Deepfakes stehen häufig im Mittelpunkt der Debatte und sind eher als mittel- bis kurzfristige Gefahren einzustufen, deren Folgen sich aller Voraussicht nach erst dann zeigen werden, wenn Deepfakes häufiger zum Einsatz kommen oder einzelne Deepfakes eine sehr große gesamtgesellschaftliche Wirkung entfalten. Theoretisch können Deepfakes beliebige Gesellschaftsbereiche betreffen, doch in der Debatte werden insbesondere die möglichen Folgen für das Medien-, Justiz-, Wissenschafts- und Wirtschaftssystem, die nationale Sicherheit, die internationalen Beziehungen und in letzter Konsequenz für das Vertrauen in die Demokratie diskutiert (Schick, 2020) (Tab. 1).

4 Regulierung von Deepfakes

Wie bei vielen neuen Technologien, ist auch die Nutzung und Verbreitung von Deepfakes einerseits in die bestehende Rechtslage eingebettet und bringt andererseits neue regulatorische Herausforderungen und damit Regulierungsbedarfe mit sich. Eine unmittelbare Regulierung von Deepfakes ist bislang weltweit nur in wenigen Staaten anzutreffen. Neben der Europäischen Union sind Regulierungs-

Tab. 1 Überblick über unterschiedliche, mit Deepfakes assoziierte Risiken (van Huijstee et al., 2021, S. 29, eigene Übersetzung)

Psychologischer Schaden	Finanzieller Schaden	Gesellschaftlicher Schaden
• Erpressung • Verleumdung • Einschüchterung • Mobbing • Rufschädigung • Untergrabung des Vertrauens • Unterdrückung der Freiheit der Meinungsäußerung	• Erpressung • Identitätsdiebstahl • Betrug (z. B. Versicherung/Zahlung) • Manipulation von Aktienkursen • Markenschädigung • Reputationsschaden	• Manipulation der Berichterstattung • Beeinträchtigung der wirtschaftlichen Stabilität • Schaden für das Justizsystem • Schädigung des Wissenschaftssystems • Erosion des Vertrauens • Schädigung der Demokratie • Manipulation von Wahlen • Beeinträchtigung internationaler Beziehungen • Beeinträchtigung der nationalen Sicherheit

bemühungen insbesondere in den Vereinigten Staaten und in China erkennbar. Im Folgenden werden zunächst die dortigen Entwicklungen skizziert. Daran schließt sich die Diskussion der einschlägigen Vorgaben auf EU-Ebene an. Schließlich werden der DSA- und KI-Regulierungsvorschlag der EU-Kommission vorgestellt und diskutiert.

4.1 Regulierungsbemühungen in den Vereinigten Staaten und in China

Die US-Reaktionen bestehen hauptsächlich aus Maßnahmen auf Bundesstaatsebene, die seit kurzem auch durch solche auf Bundesebene ergänzt wurden. Kalifornien und Texas verabschiedeten bereits 2019 Gesetze, die innerhalb eines Zeitraums im Vorfeld von Wahlen die Verbreitung manipulierter Inhalte über politische Kandidaten verbieten. Virginia verabschiedete 2019 ein Gesetz, das die Verbreitung manipulierter Inhalte unter Strafe stellt, sofern diese auf die Verleumdung, Einschüchterung oder Erpressung einer Person abzielen. Hierdurch soll vor allem die Verbreitung von nicht-einvernehmlicher Deepfake-Pornographie unterbunden werden (Feeney, 2021). Kalifornien und New York führten ein privates Klagerecht für Opfer von Deepfake-Pornographie ein, um individuelle Klagen zu vereinfachen (Ferraro & Tompros, 2020).

Auf US-Bundesebene beschränken sich die erlassenen Maßnahmen bislang auf die Systematisierung und Institutionalisierung der Sammlung relevanter Informationen über Deepfakes, mit dem Ziel der späteren Nutzung dieser für weitere fundierte Maßnahmen. Der IOGAN Act sieht etwa vor, dass die National Science Foundation Forschungsvorhaben zur Echtheitsanalyse von Deepfakes fördert und das National Institute of Standards and Technology in Kooperation mit dem Privatsektor an Möglichkeiten der Deepfake-Detektion arbeitet. Der U.S. National Defense Authorization Act 2021 adressiert zum zweiten Mal in Folge Deepfakes. Zum einen verpflichtet es das US Department of Homeland Security (DHS) dazu, Möglichkeiten zur Produktion, Erkennung und Bekämpfung von Deepfakes zu fördern. Zum anderen soll das DHS über einen Zeitraum von fünf Jahren einen jährlichen Bericht über den Einsatz gefälschter digitaler Inhalte und deren Gefahren für die Öffentlichkeit erarbeiten (Briscoe, 2021; Schapiro, 2020).

In China ist Anfang 2020 ein Gesetz in Kraft getreten, das zum einen alle Anbieter von Apps zur Produktion von Deepfake-Inhalten zur Kennzeichnung und zum anderen Plattformbetreiber dazu verpflichtet, nicht-gekennzeichnete Inhalte eigenständig zu erkennen, zu kennzeichnen oder zu entfernen, sofern es sich um unerwünschte Inhalte handelt. Schließlich soll die Gesetzesdurchsetzung mittels ergänzender Maßnahmen unterstützt werden. Eine Anmeldung bzw. Nutzung von Plattformen soll nur unter Angabe der Bürger-ID oder der Handynummer möglich sein, Plattformen sollen Beschwerdemöglichkeiten zur Meldung verdächtiger Inhalte einrichten, Audio- und Video-Plattformen sollen Standards erarbeiten und mittels eines Kreditsystems die Handlungen der Nutzenden bewerten. Darüber hinaus sollen staatliche Behörden regelmäßige Kontrollen über die Einhaltung der Gesetze durchführen, um die Verbreitung eines jeden Deepfake-Inhalts verfolgbar zu machen (Au, 2019; Chiu, 2019).

4.2 Einschlägige Vorgaben auf EU-Ebene

Die Regulierung von Deepfakes ist erst kürzlich in den Blick der EU-Politik gerückt und wird vor allem im Kontext der DSA- und KI-Verordnungsvorschläge diskutiert. Doch auch abseits der sich konkret auf Deepfakes beziehenden Regelungen spielt das Regulierungsgeflecht der EU zur Adressierung der von Deepfakes ausgehenden Risiken eine wichtige Rolle. Dieses besteht aus mitgliedstaatlichen und EU-Verfassungsnormen sowie harten und weichen Vorschriften – sowohl auf EU-Ebene als auch auf Ebene der Mitgliedstaaten. Beim Blick auf die EU-Regulierungen kann zwischen Maßnahmen unterschieden werden, die sich auf die in Deepfakes dargestellten Inhalte beziehen (die DS-GVO, das

Urheberrecht, die Richtlinie zur Bekämpfung des sexuellen Missbrauchs und der sexuellen Ausbeutung von Kindern sowie der Kinderpornografie und die Verordnung zur Bekämpfung der Verbreitung terroristischer Online-Inhalte), und solchen, die den Prozess der Zirkulation von Deepfakes regulieren. Hierzu zählen die AVMD-Richtlinie, der KI-Regulierungsvorschlag und insbesondere der DSA und die EU-Maßnahmen gegen Desinformation.

Deepfake-spezifische Gesetze, wie sie in einigen US-Bundesstaaten erlassen wurden, existieren in der EU nicht. Jedoch entfalten die EU-Regulierungen zur Inhalteregulierung in Kombination mit den in den Mitgliedstaaten bestehenden Rechten zum Schutz vor Verleumdung, Einschüchterung usw. eine ähnliche Wirkung. Vergleichbare konzertierte Maßnahmen wie in den Vereinigten Staaten hinsichtlich der Systematisierung und Institutionalisierung der Untersuchung der Folgen von Deepfakes und von neuen Detektionstechnologien hat es in der EU noch nicht gegeben. Die weitgehenden chinesischen Überwachungsmaßnahmen zur Kontrolle von Deepfakes erscheinen vor dem Hintergrund des Wertekanons der EU bzw. mit Blick auf grundlegende Menschenrechte als nicht wünschenswert, sodass über derartige Maßnahmen nicht weiter diskutiert wird (van Huijstee et al., 2021, S. 58 ff.).

Die **EU-Datenschutz-Grundverordnung** (**EU-DS-GVO**), die das Datenschutzrecht EU-weit weitestgehend harmonisiert hat, berührt Deepfakes, da Stimmfragmente oder Fotos und Videos, die zur Abbildung einer Person in einem Deepfake genutzt werden, als personenbezogenes Datum einzustufen sind. Zudem ist die DS-GVO auch auf die Entwicklung von Deepfake-Software anwendbar, weil auch zur Entwicklung solcher Software auf mit personenbezogenen Daten befüllte Foto- und Videodatenbanken zurückgegriffen wird. Als Rechtsgrundlage für die Verarbeitung personenbezogener Daten zur Erstellung eines Deepfakes kommen grundsätzlich die Einwilligung und berechtigte Interessen infrage. Letztere dürfte vor allem anwendbar sein, wenn Personen des öffentlichen Lebens unter Berufung auf das Recht auf freie Meinungsäußerung z. B. auf satirische Weise abgebildet werden. Bei Abbildung gewöhnlicher Personen, also in der Mehrzahl der Fälle, dürfte eine Einwilligung notwendig sein. Betroffene haben dann das Recht, der Verarbeitung zu widersprechen und etwa die Löschung der Videos zu verlangen. Aufgrund der Datenflut im Internet dürfte es allerdings schwierig werden, die Täter:innen zu identifizieren oder, insbesondere in grenzüberschreitenden Fällen, den Rechtsweg zu beschreiten (van Huijstee et al., 2021, S. 38 f.).

Das **Urheberrecht** in der EU basiert zwar weiterhin größtenteils auf mitgliedstaatlichem Recht, doch durch eine Reihe von Richtlinien und zwei Verordnungen ist es weitgehend harmonisiert (Margoni, 2016). Da zur Produktion

von Deepfakes häufig vorhandenes Foto- und Videomaterial verwendet wird, das in vielen Fällen urheberrechtlich geschützt sein dürfte (etwa Filmszenen), ist auch das Urheberrecht bei der Produktion und Verbreitung von Deepfakes zu beachten. Ähnlich wie im Bereich des Datenschutzrechts bieten sich hier zwei Wege: Zum einen müssen Deepfake-Produzent:innen vor der Verwendung geschützten Materials grundsätzlich die Erlaubnis der Urheberrechtsinhaber einholen. Zum anderen ermöglichen Ausnahmen die Verwendung für wissenschaftliche und künstlerische Zwecke, z. B. in Karikaturen oder Parodien (van Huijstee et al., 2021, S. 40).

Einen Hebel zur Bekämpfung von Desinformation und von Hate Speech auf Online-Videoplattformen stellt die EU-Richtlinie über audiovisuelle Mediendienste (**AVMD-Richtlinie**) dar, die 2018 in Reaktion auf Veränderungen der Medienlandschaft überarbeitet und verabschiedet wurde. Insbesondere mit Blick auf den Schutz von Minderjährigen, aber auch generell im Hinblick auf den Schutz von Medienkonsument:innen vor Volksverhetzung und Hate Speech, sieht die Richtlinie vor, dass Online-Video-Sharing-Plattformen Maßnahmen (z. B. Altersprüfung, PIN-Codes, Kennzeichnungen oder automatische Filterung) im Einklang mit der Achtung der geltenden Grundrechte und -freiheiten einführen sollen (Broughton Micova, 2020). Da sich die Regelungen der Richtlinie insbesondere auf pornographische und gewalttätige Inhalte beziehen, könnte sie geeignet sein, die Verbreitung von nicht-einvernehmlichen Deepfake-Pornos einzudämmen (van Huijstee et al., 2021, S. 42). Weil die AVMD-Richtlinie keine weiteren inhaltlichen Vorgaben macht, sondern auf die Durchsetzung des geltenden Rechts verweist, muss sich die Ergreifung von Maßnahmen in Bereichen abseits pornographischer und gewalttätiger Inhalte auf andere harmonisierte EU-Regelungen stützen. Hier kommen insbesondere die Ende 2011 in Kraft getretene Richtlinie zur Bekämpfung des sexuellen Missbrauchs und der sexuellen Ausbeutung von Kindern sowie der Kinderpornografie (Gercke, 2012) und die Verordnung zur Bekämpfung der Verbreitung terroristischer Online-Inhalte, die ab Juni 2022 anwendbar sein wird (Signorato, 2021), infrage, um Inhalte entfernen zu lassen.

4.3 EU-Maßnahmen gegen Desinformation

Den Anfangspunkt der EU-Maßnahmen gegen Desinformation bildete der Aufruf der EU-Staatschefs im März 2015 zur Erarbeitung eines Aktionsplans (European Council, 2015, S. 5). In der Folge verabschiedete das Europäische Parlament eine Entschließung zum Thema Online-Plattformen, in der die Kommission zur Vorlage

eines Gesetzesvorschlags zur Bekämpfung von sog. Fake News aufgerufen wurde (European Parliament, 2017). Die Europäische Kommission initiierte Ende 2017 eine öffentliche Konsultation zum Thema Fake News, rief Anfang 2018 eine hochrangige Expert:innengruppe zu Fake News und Online-Desinformation ins Leben und verabschiedete im April 2018 ein europäisches Konzept zur Bekämpfung von Desinformation inkl. Deepfakes im Internet. Ein zentrales Element darin war der sog. *Code of Practice on Disinformation,* der einige Wegmarken zur Selbstregulierung der Plattformbetreiber definierte (Europäische Kommission, 2018). Das bis dahin weitreichendste EU-Werkzeug, das zur Bekämpfung von Desinformation und Deepfakes verabschiedet wurde, folgte im Dezember 2018 in Form des *Aktionsplans gegen Desinformation*, der aus der Feder sowohl der Europäischen Kommission als auch des hohen Vertreters der EU für Außen- und Sicherheitspolitik gemeinsam stammt. Vor dem Hintergrund der wahrgenommenen Manipulation demokratischer Wahlen sah der Aktionsplan im Gegensatz zu den vorherigen Selbstregulierungsmaßnahmen einen stärkeren Fokus auf verbindliche Maßnahmen zur Erkennung und Bekämpfung von Desinformation vor (European Commission, 2018).

Mit dem im Dezember 2020 vorgestellten *Europäischen Aktionsplan für Demokratie* veröffentlichte die Kommission einen umfassenden Maßnahmenkatalog, der den Fokus auf drei Schwerpunkte legte: (1) die Förderung freier und gerechter Wahlen; (2) die Stärkung der Medienvielfalt und -freiheit sowie (3) die Bekämpfung von Desinformation. Hervorzuheben ist insbesondere, dass die Kommission mit dem Aktionsplan den auf Selbstregulierung fußenden Weg, den sie mit dem *Code of Practice on Disinformation* eingeschlagen hatte, weiter revidierte und verstärkt auf Elemente der Ko-Regulierung zu setzen begonnen hat. In diesem Sinne wurde insbesondere die Veröffentlichung von Leitlinien zur Überarbeitung des *Code of Practice* unter Berücksichtigung von dessen Kompatibilität mit dem DSA angekündigt (Europäische Kommission, 2020). Diese Leitlinien legte die Kommission Mitte 2021 vor. Hinsichtlich der Stärkung der Verbindlichkeit des Verhaltenskodex' sieht die Kommission die Einrichtung eines Transparenzzentrums, einer ständigen Taskforce unter Beteiligung relevanter Stakeholder sowie die Einführung von Leistungsindikatoren zur standardisierten Messung der Ergebnisse und Auswirkungen des Verhaltenskodex' vor. Zudem sollen sich die Dienstebetreiber dazu verpflichten, Empfehlungssysteme transparenter zu gestalten, leicht zugängliche Instrumente zur Meldung von Desinformation bereitzustellen, Revisionsmöglichkeiten zu eröffnen und schließlich Maßnahmen zur Erhöhung der Sichtbarkeit zuverlässiger Informationen einleiten, was auch die Interaktion mit Faktencheck-Teams einbezieht (Europäische Kommission, 2021).

4.4 Relevante EU-Rechtsreformen: Digital-Services-Act- und KI-Regulierungsvorschlag der EU-Kommission

Die zwei weitreichendsten Maßnahmen zur Adressierung der durch Deepfakes befürchteten Probleme folgten in Form zweier Verordnungsvorschläge der Europäischen Kommission. Diese werden im Folgenden skizziert und mit Blick auf Berührungspunkte zu Deepfakes diskutiert.

4.4.1 EU-Digital-Services-Act-Regulierungsvorschlag

Der im Dezember 2020 vorgelegte Verordnungsvorschlag *Digital Services Act* sieht die Harmonisierung der an Intermediäre gestellten EU-weiten Haftbarkeitsregelungen und Moderationsverpflichtungen vor.

Der DSA-Vorschlag ist eine Weiterentwicklung der *E-Commerce-Richtlinie*, die über Jahre der zentrale Baustein in der Regulierung von Online-Inhalten auf EU-Ebene war, und adressiert einige Schwachpunkte. Das Ziel der im Jahr 2000 verabschiedeten Richtlinie war es, den freien Verkehr mit Waren und Dienstleistungen im *E-Commerce* mittels EU-weit harmonisierter Regeln zu gewährleisten. Allerdings stellte die Richtlinie inhaltlich einen Minimalkompromiss dar, um die wirtschaftliche Entwicklung von Diensteanbietern, im damals noch jungen Internet, nicht durch den Erlass von übermäßigen regulatorischen Anforderungen zu hemmen. Im Hinblick auf das Thema Inhalteregulierung wurde insb. von einer Verpflichtung von Diensteanbietern zur Ex-ante-Kontrolle der über ihre Kanäle fließenden Informationen abgesehen. Inhalte sollten nur dann entfernt werden, sobald die Betreiber über deren Illegalität in Kenntnis gesetzt werden. In diesem Sinne wäre die Richtlinie auf Deepfakes grundsätzlich anwendbar und würde den Diensteanbietern vorgeben, ein Deepfake nach Kenntnis über dessen Illegalität zu löschen. Allerdings wurde in der E-Commerce-Richtlinie nicht festgelegt, ab wann ein Inhalt als illegal einzustufen ist. Die entsprechende Einstufung war und ist EU-weit stark fragmentiert und basiert auf einer Mischung aus mitgliedstaatlichen und unionalen Rechten und Gesetzen. Ungeregelt blieb auch, was genau unter „in Kenntnis gesetzt werden“ zu verstehen ist, ob bspw. Nutzermeldungen bereits als ausreichend einzustufen sind oder lediglich Gerichtsbeschlüsse akzeptiert werden. Dementsprechend wurde der rechtlichen Status Quo seit Erlass der Richtlinie als Zustand weitgehender Rechtsunsicherheit gedeutet. Zudem regelt die Richtlinie zwar, wann Dienstebetreiber von der Haftung befreit sind, aber nicht, unter welchen Bedingungen die Haftung eines Dienstebetreibers gegeben ist. Schließlich harmonisierte die Richtlinie auch nicht die Verfahrensgarantien für Widersprüche im Falle ungerechtfertigter Inhaltelöschungen, sodass Betroffene, deren

Inhalte unrechtmäßig entfernt wurden, nur in einigen Mitgliedstaaten das Recht haben, dagegen vorzugehen. Obwohl die Kommission spätestens seit 2012 im Bilde über die Schwächen der Richtlinie war, folgte sie zunächst weiterhin einem selbstregulativen Ansatz. So wurde statt des Erlasses EU-weit harmonisierter Regelungen insbesondere im Rahmen der oben diskutierten EU-Maßnahmen gegen Desinformation auf die Selbstregulierung großer Online-Plattformen gesetzt, sich der Adressierung der bekannten Schwachstellen freiwillig anzunehmen (Madiega, 2019).

Wenige Jahre später erließen einige Mitgliedstaaten angesichts der Untätigkeit der Kommission eigene Gesetze zur Adressierung der Rechtsunsicherheit im Bereich der Haftbarmachung von Intermediären. Das deutsche *NetzDG* und das französische *Gesetz gegen Falschnachrichten (Loi n° 2018–1202)* sind Ausdruck dieser Entwicklung. Die sich aufgrund des Erlasses nationaler Regulierungen abzeichnende Rechtsfragmentierung bewegte die Kommission schließlich dazu, Ende 2020 ihre Vorschläge zum Digital Services Act (DSA) und Digital Markets Act vorzulegen und damit die E-Commerce-Richtlinie zu ersetzen und mittels harmonisierter Regeln mehr Rechtssicherheit zu schaffen.

Der DSA-Vorschlag setzt den Regulierungsansatz der E-Commerce-Richtlinie fort und schreibt die Pflicht zur Löschung von illegalen Inhalten nur dann vor, sobald Dienstebetreiber Kenntnis darüber erlangen. Eine Harmonisierung dessen, was als illegaler Inhalt gilt, sieht auch der DSA-Vorschlag nicht vor und überlässt diesbezügliche Harmonisierungsbestrebungen bereichsspezifischen Regulierungen wie der DS-GVO und den anderen oben besprochenen Regulierungen.

Der **DSA** schreibt vor, dass Dienstebetreiber transparent machen müssen, welche Moderationsregeln auf ihrer Plattform gelten und welche Maßnahmen sie zu deren Durchsetzung ergreifen. Der freiwillige Einsatz von *Upload-Filtern*, also die automatisierte Filterung von Inhalten beim Hochladen, ist gemäß Kommissionsvorschlag explizit zur Durchsetzung der Moderationsregeln erlaubt. Die Betreibenden eines Koch-Forums beispielsweise dürfen, nachdem Nutzende über die Moderationsregeln und den Einsatz von Upload-Filtern informiert wurden, Inhalte automatisiert herausfiltern, die nicht den Moderationsregeln entsprechen. Eine Vorabkontrolle hochgeladener Inhalte im Hinblick auf deren Vereinbarkeit mit mitgliedstaatlichem und/oder EU-Recht sieht der DSA-Vorschlag hingegen nicht vor (Reda, 2021).

Darüber hinaus werden Betreiber, die ein *Notice-and-Takedown*-Verfahren anwenden, dazu verpflichtet, ein System zu schaffen, mit dem illegale Inhalte EU-weit harmonisiert gemeldet werden können. Dadurch soll es Betroffenen erleichtert werden, die gegen sie gerichteten offensichtlich rechtswidrigen Inhalte

zu melden und entfernen zu lassen. Dadurch, dass der Vorschlag dies nicht explizit ausschließt, kann eine automatisierte Löschung und Sperrung gemeldeter Inhalte nicht ausgeschlossen werden (ebd.). Wenn ein Verdacht auf eine schwere Straftat vorliegt, die eine Bedrohung für das Leben oder die Sicherheit von Personen darstellt, sollen Betreiber diesen Verdacht zudem an die zuständigen Strafverfolgungsbehörden weitergeben müssen.

Weil im DSA-Vorschlag Plattformen ab einer bestimmten Größe als quasi-öffentlicher Diskursraum betrachtet werden, auf dem die Ausübung des Rechts auf freie Meinungsäußerung zu schützen ist, sieht der DSA-Vorschlag zudem die Einrichtung eines Verfahrens vor, mit dem Betroffene gegen eine Sperrung vorgehen können. Daneben verpflichtet das Gesetz alle Betreiber dazu, mehr Transparenz über die erfolgten Sperrungen herzustellen, indem sie eine Datenbank aufbauen, die u. a. Informationen über den Grund der Sperrung und den jeweiligen Beschwerdeführer beinhalten (European Commission, 2020; Schünemann, 2021).

4.4.1.1 Der Digital Services Act und Deepfakes

Da sozialen Online-Netzwerken eine zentrale Rolle bei der **Verbreitung von Deepfakes** zukommt, betreffen die Vorgaben des DSA Deepfakes unmittelbar. So könnten die harmonisierten *Notice-and-Takedown*-Regeln eine wirksamere Löschung unrechtmäßiger Deepfakes ermöglichen. Zugleich wären Upload-Filter zur Rechtsdurchsetzung ausgeschlossen. Befürchtet wird jedoch, dass die implizite Erlaubnis zum Einsatz von Upload-Filtern zur Durchsetzung der Moderationsregeln eine de facto Einschränkung des Rechts auf freie Meinungsäußerung bewirken könnte, wenn sich Intermediäre selbst mit nationalem bzw. EU-Recht konforme Moderationsregeln auferlegen, um möglicherweise entstehende Mühen bei der Inhaltelöschung zu vermeiden. In ähnlicher Weise wäre auch denkbar, dass die Betreiber trotz der Transparenzvorgaben in Bezug auf gemeldete sowie gelöschte Inhalte (und trotz der Einspruchsmöglichkeiten, die für Betroffene von Sperrungen und Löschungen vorgesehen sind) ein *Overblocking* mittels automatisierter Sperrungen und Löschungen betreiben könnten. Die Wiederherstellung derartiger unrechtmäßig entfernter Inhalte wäre erst nach langwierigen Beschwerdezyklen möglich, die viele Betroffene wohl nicht auf sich nehmen würden, sodass eine Beeinträchtigung des Rechts auf freie Meinungsäußerung befürchtet wird (Buiten, 2021, S. 24 f.).

Sofern ohnehin davon auszugehen ist, dass Dienstebetreiber Upload-Filter zur Durchsetzung ihrer Moderationsregeln einsetzen werden, könnten sie auch dazu verpflichtet werden, mittels Filter-Einsatzes die Authentizität sowohl hochgeladener Inhalte als auch der Accounts, die entsprechende Inhalte teilen,

zu erkennen und Gegenmaßnahmen zu ergreifen. Etwa in Form einer Kennzeichnung nicht-authentischer Inhalte oder der Unkenntlichmachung der in nicht-authentischen Videos dargestellten Gesichter oder Stimmen (van Huijstee et al., 2021, S. 62 f.).

Schließlich sei erwähnt, dass für eine angemessene Adressierung der aus Deepfakes resultierenden Herausforderungen auch Schritte notwendig wären, die über den Anwendungsbereich des DSA hinausgehen. Darunter fallen etwa mitgliedstaatliche Maßnahmen zur Unterstützung der Opfer von Deepfakes und europaweite Maßnahmen zur Aufklärung der von Deepfakes ausgehenden Gefahren (van Huijstee et al., 2021, S. 62 ff.).

4.4.2 EU-KI-Regulierungsvorschlag

Der KI-Regulierungsvorschlag der Kommission, der im April 2021 veröffentlicht wurde, betrifft das Thema Deepfakes ebenfalls in mehreren Hinsichten (European Commission, 2021).

Grundsätzlich verfolgt der Vorschlag das Ziel, die vertrauenswürdige und sichere Anwendung Künstlicher Intelligenz unter Einhaltung der EU-Werte und -Grundrechte zu ermöglichen. Zu diesem Zweck werden harmonisierte Regeln für die Entwicklung, Verbreitung und Nutzung von KI-Systemen vorgeschlagen. Dabei verfolgt der Kommissionsvorschlag einen risikobasierten Ansatz und unterteilt KI-Systeme in vier Risikokategorien: 1. Unannehmbares Risiko; 2. Hohes Risiko; 3. Geringes Risiko und 4. Minimales Risiko. Während Systeme, die ein unannehmbares Risiko mit sich bringen, verboten werden sollen, werden für Hochrisiko-Systeme strenge Vorgaben, für Systeme mit geringem Risiko vor allem Transparenzvorgaben und für solche mit einem minimalen Risiko keine Vorgaben vorgeschlagen (Geminn, 2021).

4.4.2.1 EU-KI-Regulierungsvorschlag und Deepfakes

Deepfakes werden im Vorschlag explizit erwähnt und gemäß Kommission unter der Kategorie der Systeme mit geringem Risiko gefasst. Nach Artikel 52 (3) sollen Nutzende eines KI-Systems, das zur Produktion von Deepfakes dient, offenlegen, dass die dargestellten Inhalte künstlich erzeugt oder manipuliert wurden, womit eine gesetzliche EU-weite Kennzeichnungspflicht für Deepfakes eingeführt würde.[4] Zugleich wird im Erwägungsgrund 38 und im Annex III

[4] Ausnahmen sollen für den Bereich der Strafverfolgung, für die Ausübung des Rechts auf freie Meinungsäußerung sowie die Kunst- und Wissenschaftsfreiheit gelten (vgl. Art. 52 (3)).

(European Commission, 2021, S. 4) geregelt, dass die Nutzung von Deepfake-Detektionstechnologien durch Strafverfolgungsbehörden unter die Hochrisiko-Kategorie und die damit verbundenen strengen Vorgaben fällt.

Doch sind mit dem Kommissionsvorschlag auch einige Probleme verbunden. Diese Schwierigkeiten ergeben sich einerseits aus dem Anwendungsbereich und andererseits aus dem unzureichenden Strafbezug. Denn die Kennzeichnungspflicht aus Artikel 52 (3) betrifft lediglich die Nutzenden eines KI-Systems, nicht aber die Hersteller und Anbieter der Systeme. Schaut man sich die geläufigen FaceSwap-Apps an, mit denen einfache Deepfakes erstellt werden können, sind entsprechende Markierungen in Form der Hersteller-Logos stets softwareseitig in die produzierten Videos eingebettet. Im Falle, dass die Hersteller und Anbieter dies nicht von sich aus anbieten, müssten allerdings die Nutzenden selbst entsprechende Kennzeichnungen in die Deepfake-Inhalte per Wasserzeichen einbetten oder dies bei Veröffentlichung der Inhalte in einem Begleitkommentar kenntlich machen. Dies wäre freilich nur dann ein Problem, wenn die Anbieter und Hersteller nicht mehr selbständig entsprechende Kennzeichnungen vornehmen. Eine an die Anbieter und Hersteller gerichtete Verpflichtung zur Kennzeichnung würde hier aber Gewissheit schaffen. Zu einem echten Problem wird die unzureichende Kennzeichnungspflicht schließlich erst dann, wenn es um die Bekämpfung ungewollter Deepfake-Inhalte geht. Denn Deepfakes, die etwa mit dem Ziel der Desinformation oder Verunglimpfung verbreitet werden, sollen ja gerade nicht als Fälschung oder Manipulation erkannt werden, sodass weder die Hersteller und Anbieter noch die Nutzenden der entsprechenden KI-Systemen die Inhalte freiwillig kennzeichnen würden. In diesen Fällen entstünde mit der fehlenden Kennzeichnungspflicht für Anbieter und Hersteller schlicht eine Gesetzeslücke. Das Herstellen und Anbieten von Deepfake-Produktionssoftware, die bewusst keine Kennzeichnung beinhaltet, sodass Täuschungen Vorschub geleistet wird, wäre nicht strafbar[5] (van Huijstee et al., 2021, S. 44 f.). Unklar ist überdies, wie die zur Rechtsdurchsetzung vorgesehenen Aufsichtsbehörden mutmaßlich nicht-gekennzeichnete Deepfakes, die durch Nutzende verbreitet wurden, erkennen können sollen (Veale & Borgesius, 2021, S. 20). Problematisch ist auch, dass die Verbreitung von KI-basierten Deepfake-Detektionstechnologien

[5] Erwähnt sei, dass einige Anbieter und Hersteller entsprechender KI-Systeme im Falle einer Strafbarkeit der Nicht-Implementierung einer noch in das KI-Gesetz zu integrierenden Kennzeichnungspflicht wahrscheinlich andere Strafumgehungsmöglichkeiten nutzen und ihre Dienste bspw. aus dem EU-Ausland heraus anbieten würden, sodass für diese Herausforderung wieder neue Lösungen gefunden werden müssten.

im Vorschlag nicht geregelt wird, obwohl dadurch die Verbreitung und Entwicklung kriminell motivierter Deepfake-Technologien gebremst werden könnte. Zudem fallen einige der zur Adressierung der Herausforderungen von Deepfakes diskutierten Maßnahmen gänzlich aus dem Anwendungsbereich des Verordnungsvorschlags. Zur Entwicklung von verbesserten Detektionstechnologien wären z. B. Schritte im Rahmen des EU-Forschungsrahmenprogramms notwendig (van Huijstee et al., 2021, S. 59 ff.).

5 Schlussfolgerungen

Der Aufsatz hat einen Überblick zum Stand der Regulierung von Deepfakes in der EU geliefert. Mittels einer einführenden konzeptionellen Einordnung wurde zunächst gezeigt, dass Deepfakes als Querschnittsphänomen mit Diskursen zu Desinformation, Fehlinformation, zu schädigenden Informationen und Hate Speech verbunden sind. Dies zeigte sich auch bei der Diskussion der mit Deepfakes assoziierten Chancen und Risiken: Einerseits stehen vor allem die von der Technologie ausgehenden Risiken im Mittelpunkt der gesellschaftspolitischen Debatte. Andererseits wird mittels einer Typisierung der Risiken vor allem jenen eine größere Bedeutung zugeschrieben, die sich auf mögliche institutionelle und gesellschaftliche Schäden beziehen. Diese Risiken stellen mögliche Entwicklungsszenarien dar, deren Schäden sich meist noch nicht entfaltet haben. Durchaus reale Auswirkungen haben Deepfakes hingegen schon heute insbesondere im individuellen Bereich. Die überwältigende Mehrheit aller Deepfake-Inhalte bezieht sich auf Deepfake-Pornographie, die zudem fast ausschließlich Frauen betrifft.

Die Diskussion zur Regulierung von Deepfakes hat gezeigt, dass Deepfakes auch in anderen Teilen der Welt in den Blick des Gesetzgebers rücken. Während die Vereinigten Staaten eine zurückhaltende Regulierungspraxis betreiben und bundesweite Regulierungen sich vor allem auf die kontinuierliche Lagebewertung und technologische Hoheitsbestrebungen fokussieren, reihen sich die Maßnahmen der Volksrepublik in die Politik der weitgehenden Überwachung aller Internetinhalte ein.

Demgegenüber bauen die EU-Maßnahmen vorwiegend auf zwei Pfeilern. Zum einen werden mittels DSA- und KI-Regulierungsvorschlag konkret auf Deepfake-Technologien bezogene Maßnahmen diskutiert, mittels derer die Zirkulation und Wahrnehmung von Deepfakes reguliert werden soll. Der KI-Regulierungsvorschlag setzt hier vor allem auf die Schaffung von Transparenz. Der DSA-Vorschlag soll schließlich EU-weit harmonisierte Regeln festlegen. Diese berühren

den gesamten Zirkulationsprozess eines Deepfakes von der Meldung bis zur Entfernung von Inhalten. Mittels DSA sollen sowohl die Rechte von Personen, die einen Inhalt melden, als auch jene der von einer Entfernung ihres Inhalts Betroffenen miteinander in Einklang gebracht werden.

Beide Regulierungsvorhaben stellen zwei zentrale Grundpfeiler der EU-Digitalpolitik dar, die das Feld der Plattformregulierung und des Einsatzes Künstlicher Intelligenz auf Jahre definieren werden. Dementsprechend wird um beide Vorschläge politisch intensiv gerungen und beide Verordnungen werden, sofern der Gesetzgebungsprozess erfolgreich abgeschlossen wird, am Ende einen großen inhaltlichen Wandel durchlaufen haben. Fraglich ist dabei, inwiefern sich die Regulierungsvorschläge der Einhegung der durch Deepfakes befürchteten Gefahren annehmen werden, und noch fraglicher ist, ob die dann eventuell erlassenen Maßnahmen ein sinnvolles Werkzeug darstellen werden. Schließlich basiert ein Großteil der Forschung zu den möglichen Folgen von Deepfakes auf Annahmen. Klar ist jedenfalls, dass neben den in diesem Aufsatz erläuterten, weitere Bemühungen erforderlich sein werden, um die befürchteten Folgen von Deepfakes einzuhegen. Dazu zählen sowohl weitere Bemühungen auf EU-Ebene und insbesondere auch mitgliedstaatliche Maßnahmen.

Insofern liefert der vorliegende Beitrag mit Blick auf die Rolle der DSA- und KI-Verordnungsvorschläge bei der Regulierung von Deepfakes einen Schnappschuss. Er macht aber zugleich auch deutlich, dass die Zeiten überholt sind, in denen digitalpolitische Themen und als Gefährdung wahrgenommene technologische Entwicklungen erst mit Verzögerung in den Blick des Gesetzgebers rückten.

Danksagung Die diesem Beitrag zu Grunde liegenden Arbeiten basieren auf einer Studie, die im Auftrag des Gremiums zur wissenschaftlich-technischen Folgenabschätzung (STOA) und unter der Leitung des Referats „Wissenschaftliche Vorausschau“ der Direktion „Folgenabschätzungen und europäischer Mehrwert“ innerhalb der Generaldirektion „Wissenschaftlicher Dienst“ (EPRS) des Generalsekretariats des Europäischen Parlaments gefördert wurde.

Literatur

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. Zugegriffen: 7. Dez. 2020.

Au, L. (2019). China targets ‚deepfake‘ content with new regulation. Technode. https://technode.com/2019/12/03/china-targets-deepfake-content-with-new-regulation/Zuletzt. Zugegriffen: 6. Apr. 2021.

Ayyub, R. (2018). I was the victim of a deepfake porn plot intended to silence me. HuffPost. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316. Zugegriffen: 4. Apr. 2021.

Bateman, J. (2020). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace. https://carnegieendowment.org/files/Bateman_FinCyber_Deepfakes_final.pdf. Zugegriffen: 7. Dez. 2020.

Briscoe, S. (2021). U.S. laws address deepfakes. Security management. http://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/Zuletzt. Zugegriffen: 6. Apr. 2021.

Broughton Micova, S. (2020). The audiovisual media services directive: Balancing liberalisation and protection (Draft). SSRN. https://papers.ssrn.com/abstract=3586149. Zugegriffen: 11. Nov. 2021.

Buiten, M. (2021). The digital services act: From intermediary liability to platform regulation. SSRN. https://papers.ssrn.com/abstract=3876328. Zugegriffen: 12. Okt. 2021.

BuzzFeed. (2018). You won't believe what obama says in this video . Twitter. https://twitter.com/buzzfeed/status/986257991799222272. Zugegriffen: 28. Okt. 2021.

Centre for Data Ethics and Innovation. (2019). Snapshot paper – Deepfakes and audiovisual disinformation. GOV.UK. https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation. Zugegriffen: 6. Aug. 2021.

Channel 4. (2020). Deepfake Queen: 2020 Alternative christmas message. YouTube. https://www.youtube.com/watch?v=IvY-Abd2FfM. Zugegriffen: 29. Okt 2021.

Chesney, R., & Keats Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. SSRN. https://www.ssrn.com/abstract=3213954. Zugegriffen: 7. Dez. 2020.

Chiu, K. (2019). China announces new rules to tackle deepfake videos. abacus. https://www.scmp.com/abacus/news-bites/article/3040033/china-announces-new-rules-tackle-deepfake-videos. Zugegriffen: 19. Okt. 2020.

Collins, A. (2019). Forged authenticity: Governing deepfake risks. EPFL. https://www.epfl.ch/research/domains/irgc/specific-risk-domains/projects-cybersecurity/forging-authenticity-governing-deepfake-risks/Zuletzt. Zugegriffen: 16. März 2021.

Compton, S. (2021). More and more women are facing the scary reality of deepfakes. Vogue. https://www.vogue.com/article/scary-reality-of-deepfakes-online-abuse. Zugegriffen: 16. März 2021.

Curry, D. (2021). Most Popular Apps (2021). Business of Apps. https://www.businessofapps.com/data/most-popular-apps/Zuletzt. Zugegriffen: 29. Okt. 2021.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (Microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics, 26*, 69–91. https://doi.org/10.1177/1940161220944364

Eisenegger, M. (2021). Dritter, digitaler Strukturwandel der Öffentlichkeit als Folge der Plattformisierung. In M. Eisenegger, M. Prinzing, P. Ettinger, & R. Blum (Hrsg.), *Digitaler Strukturwandel der Öffentlichkeit* (S. 17–39). Springer VS.

Europäische Kommission. (2018). Mitteilung Der Kommission An Das Europäische Parlament, Den Rat, Den Europäischen Wirtschafts- Und Sozialausschuss Und Den Ausschuss Der Regionen Bekämpfung von Desinformation im Internet: Ein europäisches

Konzept. EUR-Lex. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52018DC0236. Zugegriffen: 16. Sept. 2018.

Europäische Kommission. (2020). Europäischer Aktionsplan für Demokratie. Europäische Kommission. https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_de. Zugegriffen: 16. Sept. 2021.

Europäische Kommission. (2021). Kommission legt Leitlinien zur Stärkung des Verhaltenskodex für den Bereich der Desinformation vor. Europäische Kommission. https://ec.europa.eu/commission/presscorner/detail/de/ip_21_2585. Zugegriffen: 16. Sept. 2021.

European Commission. (2020). Proposal for a regulation of the European parliament and of the council on a single market for digital services (Digital Services Act) and amending directive 2000/31/EC. EUR-Lex. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN Zugegriffen: 16. Sept. 2021.

European Commission. (2021). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Zugegriffen: 16. Sept. 2021.

European Council. (2015). European Council meeting (19 and 20 March 2015) – Conclusions. European Council. https://www.consilium.europa.eu/media/21888/european-council-conclusions-19-20-march-2015-en.pdf. Zugegriffen: 16 Sept. 2018.

European Parliament. (2017). European parliament resolution of 15 June 2017 on online platforms and the digital single market (2016/2276(INI)). EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0272. Zugegriffen: 16. Sept. 2018.

Feeney, M. (2021). Deepfake laws risk creating more problems than they solve. Regulatory transparency project. https://regproject.org/wp-content/uploads/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf. Zugegriffen: 31. März 2021.

Ferraro, M. F., & Tompros, L. W. (2020). New York's right to publicity and deepfakes law breaks new ground. Wilmerhale. https://www.wilmerhale.com/-/media/files/shared_content/editorial/publications/wh_publications/client_alert_pdfs/20201217-new-yorks-right-to-publicity-and-deepfakes-law-breaks-new-ground.pdf. Zugegriffen: 31 März 2021.

Geminn, C. (2021). Die Regulierung Künstlicher Intelligenz: Anmerkungen zum Entwurf eines Artificial Intelligence Act. *Zeitschrift Datenschutz, 7*, 354–359.

Gercke, M. (2012). Die EU Richtlinie zur Bekämpfung von Kinderpornographie. *Computer und Recht, 28*, 520–525.

Högden, B., Krämer, N., Meinert, J., & Schaewitz, L. (2020). Wirkung und Bekämpfung von Desinformation aus medienpsychologischer Sicht. In M. Steinebach, K. Bader, L. Rinsdorf, N. Krämer, & A. Roßnagel (Hrsg.), *Desinformation aufdecken und bekämpfen: Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität* (S. 77–99). Nomos.

Jiang, L. u. a. (2021). DeeperForensics challenge 2020 on Real-world face forgery detection: Methods and results. Cornell University. http://arxiv.org/abs/2102.09471. Zugegriffen: 5. Apr. 2021.

Johann, M., & Wagner, J. (2020). Neue Debatte, altes Dilemma? Die Herausforderungen des Phänomens »Fake News« für die Unternehmenskommunikation. In R. Hohlfeld,

M. Harnischmacher, E. Heinke, L. S. Lehner, & M. Sengl (Hrsg.), *Fake News und Desinformation Herausforderungen für die vernetzte Gesellschaft und die empirische Forschung* (S. 99–116). Nomos.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: trick or treat? *Business Horizons, 63*, 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Madiega, T. (2019). Reform of the EU liability regime for online intermediaries: Background on the forthcoming digital services act: In-depth analysis. Publications Office of the European Union. https://data.europa.eu/doi/10.2861/08522 Zugegriffen: 25. Jan. 2021.

Margoni, T. (2016). The harmonisation of EU copyright law: The originality standard. In M. Perry (Hrsg.), *Global governance of intellectual property in the 21st century: reflecting policy through change* (S. 85–105). Springer International Publishing.

Minsky, C. (2021). 'Deepfake' videos: To believe or not believe? Financial Times. https://www.ft.com/content/803767b7-2076-41e2-a587-1f13c77d1675. Zugegriffen: 29. Juni 2021.

Paris, B., & Donovan, J. (2019). Deepfakes and cheap fakes. *Data & Society* 50.

Reda, J. (2021). Edit policy: Der Digital Services Act steht für einen Sinneswandel in Brüssel. Netzpolitik.org. https://netzpolitik.org/2021/edit-policy-der-digital-services-act-steht-fuer-einen-sinneswandel-in-bruessel/Zuletzt. Zugegriffen: 15. Febr. 2021.

Rudnicki, K., & Steiger, S. (2020). Online hate speech: Introduction into motivational causes, effects and regulatory contexts. Media Diversity Institute. https://www.media-diversity.org/wp-content/uploads/2020/09/DeTact_Online-Hate-Speech.pdf. Zugegriffen: 15. Juni 2021.

Schapiro, Z. (2020). DEEP FAKES accountability act: Overbroad and ineffective. IPTF. http://bciptf.org/2020/04/deepfakes-accountability-act/Zuletzt. Zugegriffen: 6. Apr. 2021.

Schick, N. (2020). *Deep fakes and the infocalypse: What you urgently need to Know*. Monoray.

Schünemann, W. J. (2021). New horizontal rules for online platforms across Europe: A comment on the commission's proposal for a digital services act for DTCT partners and upstanders. Drct detect then act. https://dtct.eu/wp-content/uploads/2021/02/DTCT-TR1-DSA.pdf. Zugegriffen: 15. Juni 2021.

Signorato, S. (2021). Combating terrorism on the internet to protect the right to life. The regulation (EU) 2021/784 on adressing the dissemination of terrorist content online. In P. Czech, L. Heschl, K. Lukas, M. Nowak, & G. Oberleitner (Hrsg.), *Yearbook: Human Rights Protection. Right to life* (S. 403–408). Vojvodina Provincial authorities Common Affairs Department.

Sokolov, S. S., Alimov, O. M., Tyapkin, D. A., Katorin, Y. F., & Moiseev, A. I. (2020). Modern social engineering voice cloning technologies. *2020 IEEE conference of russian young researchers in electrical and electronic engineering (EIConRus)* 513–16.

Steinebach, M., Bader, K., Rinsdorf, L., Krämer, N., & Roßnagel, A. (Hrsg). (2020). *Desinformation aufdecken und bekämpfen: Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität*. Nomos.

Trend Micro Research, UNICRI, und EC3. (2020). Malicious uses and abuses of artificial intelligence. UNICRI United Nations Interregional Crime and Justice Research

Institute. https://unicri.it/sites/default/files/2020-11/AI%20MLC.pdf. Zugegriffen: 7. Dez. 2020.
van Huijstee, M., van Boheemen, P., Das, D., Nierling, L., Jahnel, J., Karaboga, M., Fatun, M., Kool, L., & Gerritsen, J. (2021). Tackling deepfakes in European policy. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf. Zugegriffen: 5. Okt. 2021.
Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU artificial intelligence Act. SocArXiv Papers. https://osf.io/38p5f. Zugegriffen: 3. Aug. 2021.
Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c. Zugegriffen: 28. Okt. 2021.