Teil IV
Künstliche Intelligenz, Desinformation und Deepfakes

Das Phänomen Deepfakes. Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung[1]

Anna Louban, Milan Tahraoui, Hartmut Aden, Jan Fährmann, Christian Krätzer und Jana Dittmann

Zusammenfassung

Bildern und Videos kommt im politischen Diskurs eine zunehmende Bedeutung zu. Auch politische Desinformation wird vermehrt in Form videographischer Inhalte transportiert. *Deepfakes* sind durch Methoden Künstlicher Intelligenz (KI) generierte oder manipulierte Bilder, Audios und Videos. Dieser Beitrag fragt interdisziplinär aus den Perspektiven der Rechts- und Politikwissenschaft sowie der Informatik nach den Risiken für politische Entscheidungsprozesse, zu denen *Deepfakes* und ihre Nutzung für politische Desinformation führen können. Darauf basierend präsentiert der Beitrag Ansätze aus dem multidisziplinär ausgerichteten Forschungsprojekt *FAKE-ID* zur Erforschung KI-basierter Deepfake-Detektoren.

1 Dieser Beitrag basiert auf dem Forschungsprojekt *FAKE-ID: Videoanalyse mit Hilfe künstlicher Intelligenz zur Detektion von falschen und manipulierten Identitäten*, gefördert vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen der Bekanntmachung *Künstliche Intelligenz in der zivilen Sicherheitsforschung*. Das Projektkonsortium erarbeitet Kriterien, anhand derer Fälschungen von KI-manipulierten Bildern und Videodatenströmen identifiziert und klassifiziert werden können (FKZ: HWR/FÖPS Berlin 13N15737, OVGU 13N15736). Den Kern der angestrebten Detektionslösung im Teilprojekt der OVGU bildet eine KI-generierte Risiko- und Verdachtslandkarte, die Authentizitätsindizien beziehungsweise Verdachtselemente in Bild- und Videomaterial visuell aufbereitet und Anwender:innen bei der Entscheidungsfindung unterstützt. Neben dem Forschungsinstitut für Öffentliche und Private Sicherheit (FÖPS Berlin) der Hochschule für Wirtschaft und Recht Berlin (HWR) und der Otto-von-Guericke-Universität Magdeburg (OVGU) zählen die Bundesdruckerei (Konsortialführung), das Fraunhofer Heinrich Hertz Institut (HHI) und die BioID GmbH zu den forschenden Konsortialpartner:innen.

1. Einleitung

Mit Künstlicher Intelligenz (KI) können Medien wie Bilder, Audios und Videos so verändert werden, dass Betrachter:innen die auf diese Weise entstandenen *Deepfakes* nicht ohne Weiteres als eine Manipulation erkennen können. Wie alle technischen Entwicklungen birgt auch KI-generiertes Bild-, Audio- und Videomaterial sowohl Potentiale für neue nützliche Anwendungen (z. B. für Unterhaltung, Kunst und Medizin)[2] als auch Risiken und Gefahren, die im politischen Kontext bereits beobachtet werden können. Dass der demokratische Diskurs durch manipuliertes oder künstlich generiertes, aber echt wirkendes Bild-, Audio-, und Videomaterial beeinflusst werden kann, zeigen bereits Erfahrungen aus der Vergangenheit. Manipulierte Bilder wurden zu propagandistischen Zwecken[3] bis hin zur Rechtfertigung für militärische Auseinandersetzungen verwendet.[4] In Anbetracht dieser historischen Erfahrungen und der stetig wachsenden Leistungsfähigkeit KI-basierter Anwendungen sind die von *Deepfakes* ausgehenden Risiken für die Funktionsfähigkeit demokratischer Staaten, etwa im Kontext von Wahlen, als bedeutend einzustufen, etwa wenn der Wahlerfolg politischer Gegner:innen durch gefälschte oder manipulierte Medien gezielt beeinträchtigt wird.

Im Wissen darum, dass Menschen visuellen Darstellungen einen besonders starken Wirklichkeitsbezug beimessen,[5] wird im Rahmen der politischen Kommunikation zunehmend mit Bildern und kurzen Videoclips gearbeitet.[6] Insbesondere die Fähigkeit, starke Emotionen bei Rezipient:innen hervorzurufen,[7] gibt Bildern den Vorzug vor primär textbasierten Mitteln der Massenkommunikation.[8] Durch die Digitalisierung und die zunehmende Verbreitung qualitativ hochwertiger Aufnahme- und Wiedergabegeräte steht immer mehr Bildmaterial zur Verfügung,[9] das innerhalb kurzer Zeit über digitale Plattformen weltweit verbreitet werden kann. Im Zuge dieser Entwicklungen können Bildinhalte kaum noch ohne tech-

2 Siehe z. B.: *European Parliamentary Research Service*, Tackling deepfakes in European policy, 2021, 28-29.
3 *Lüthe*, in: Liebert/Metten (Hrsg.), Mit Bildern lügen, 2007, 50.
4 *Hömberg / Karasek*, Communicatio Socialis 2008, 276 f.
5 *Gerth*, IMAGE 2018, 14 (27), 5; Siehe Diskussionen unter: *Steding*, Ein Bild lügt mehr als tausend Worte, o.J.
6 *Hömberg* in: Hohlfeld u. a. (Hrsg.), Fake News und Desinformation, 2020, 83.
7 *Bessette-Symons*, Memory 2018, 171.
8 *Isermann / Knieper*, in: Schicha / Brosda (Hrsg.), Handbuch Medienethik, 2017, 304.
9 Vgl. *Fährmann,* MMR 2020, 228.

nische Unterstützung auf ihre Echtheit überprüft werden. Der Umstand, dass Manipulationen von online verfügbaren Inhalten zunehmend auf dem Einsatz von Künstlicher Intelligenz (KI) basieren, legt auch den Einsatz von KI bei der Entwicklung von Gegenmaßnahmen nahe.

Vor diesem Hintergrund geht der vorliegende Beitrag zunächst der Frage nach, wie *Deepfakes* politische Diskurse beeinflussen können. Da sowohl die freie politische Meinungsäußerung als auch die unbeeinflusste Meinungsbildung zentral für die Ausgestaltung demokratischer Prozesse sind, bedarf das Phänomen *Deepfakes* einer differenzierten Betrachtung. In einem zweiten Schritt setzt sich der Beitrag mit rechtlichen Strategien für den Umgang mit *Deepfakes* auseinander. Im dritten Schritt wird das interdisziplinäre Konsortialprojekt *FAKE-ID* vorgestellt, das unter anderem KI-basierte *Deepfake*-Detektionstools erforscht.

2. *Deepfakes – ein KI-Phänomen mit vielfältigen Einsatzmöglichkeiten*

Der Begriff *Deepfake*[10] wird aus den beiden Begriffen *Deep Learning* und *Fake* gebildet und beschreibt einen Teilbereich der Künstlichen Intelligenz (KI), der auf die Erstellung einer Fälschung mittels der Methode des *Deep Learning* abzielt. Er wird oftmals für ein Video verwendet, das mithilfe von *Deep-Learning*-Methoden bearbeitet wurde, um die Person im Originalvideo partiell (hinsichtlich Gesicht, Körper, Gestik, Sprache, o. ä.) durch eine andere Person auf eine Art zu ersetzen, die in der Rezeption den Eindruck einer unverfälschten und somit glaubwürdigen Darstellung erzeugt. In vielen Fällen handelt es sich bei den vermeintlichen Protagonist:innen um Personen des öffentlichen Lebens.[11]

Der Transfer des Begriffs *Deepfake*[12] aus der Techniksprache in den öffentlichen Diskurs lässt sich auf das Jahr 2017 datieren, als eine(r) der an-

10 Die Arbeiten an der Otto-von-Guericke-Universität Magdeburg (OVGU) zur Definition von Deepfakes sind zusammen mit Dennis Siegel und Stefan Seidlitz erfolgt. Wir danken beiden für die konstruktive Diskussion.

11 Siehe zum Beispiel: *Webster,* Words We're Watching, 'Deepfake'.

12 Die Bedeutung des Begriffs *Deepfake* formuliert die *Duden-Redaktion* (2022) als „(z. B. in krimineller oder satirischer Absicht) mithilfe künstlicher Intelligenz erzeugte beziehungsweise manipulierte Bild- oder Tondatei". Seit dem Jahr 2021 wird der Anglizismus *Deepfake* im Online-Dudenwörterbuch geführt, siehe: *Weiffen*, Der Anglizismus des Jahres 2019 lautet ‚for future'. Im gleichen Jahr kürte die Anglizismus-Jury den Begriff *Deepfake* zum drittwichtigsten Ausdruck, der aus der englischen Sprache in die deutsche integriert worden ist. Siehe: *Stefanowitsch* u. a., Anglizismus des Jahres 2019.

onymen Nutzer:innen der Onlineplattform *Reddit* unter dem Pseudonym *deepfakes* Videomaterial veröffentlichte, das unter Einsatz von KI Gesichter von aus der Film- und Musikindustrie bekannten Frauen auf die Körper von Pornodarstellerinnen montierte.[13] Das auf diese Weise erzeugte Filmmaterial wirkte – zumindest auf den ersten Blick – glaubwürdig. Bereits kurze Zeit nach diesem Ereignis konnte man beobachten, wie die Nachfrage nach (gefälschten) pornographischen Inhalten zur vermehrten Verbreitung von *Deepfakes* führte.[14] Frei verfügbare Software und Bedienungsanleitungen für die Generierung von Bildern und Videos pornographischen Inhalts trugen dazu bei, dass 96 % der insgesamt 14.678 im Jahr 2019 von Ajder et al. untersuchten *Deepfakes* der Kategorie „non-consensual deepfake pornography" zugeordnet werden konnten.[15]

In den darauffolgenden Jahren haben die Formate digitaler Fälschungen und Manipulationen merklich an Variation zugenommen. Insofern erscheint es sinnvoll und notwendig, die Definition von *Deepfakes* auszuweiten: *Deepfakes* sollen demnach als generierte, potentiell glaubwürdige Medieninhalte (Bilder, Videos, Texte,[16] Audiodaten, etc.) verstanden werden, die durch die teilweise Verfälschung von bestehenden medialen Inhalten (z. B. Videomaterial) zumeist mittels eines neuronalen Netzwerkes produziert werden.[17]

Auch die Einsatzgebiete für KI-basierte Manipulationen von Bild, Video- und Audioinhalten erweiterten und diversifizierten sich. Insbesondere die Bereiche Kommerz, Unterhaltung und Medizin profitieren von den Möglichkeiten der gezielten Veränderung digitaler Daten durch KI. Diverse weitere Einsatzbereiche für *Deepfakes* werden in der Fachliteratur diskutiert, etwa die Beseitigung von Sprachbarrieren zur Verbesserung kulturübergreifender Verbreitung von Videoinhalten oder zur direkten politischen Ansprache. Andere Einsatzmöglichkeiten bieten sich in der Bildbearbeitung in der Filmindustrie, der Erstellung personalisierter Medien, der Produktion von KI-Werbemodellen unter Verwendung von

13 *Panholzer,* Deepfakes wurden durch Pornografie bekannt.

14 Siehe z. B.: *Hardford*, Does pornography still drive the Internet?, 2019; *Waddel*, How Porn Leads People to Upgrade Their Tech, 2016.

15 *Ajder u. a.*, The State of Deepfakes, 2019.

16 Vgl. Brando Benifei, Änderungsantrag 753 zu Artikel 52 – Absatz 3– Einleitung, der vorschlägt „***Text***-[…]inhalte" (Herv. i. O.) in die Aufzählung der Medien aufzunehmen, die durch Künstliche Intelligenz so „erzeugt oder manipuliert" werden können, dass ein „Deepfake" entsteht. https://www.europarl.europa.eu/doceo/document/JURI-AM-730042_DE.pdf
Siehe auch: https://mixed.de/facebook-zeigt-deepfake-text-ki-und-warnt-davor/.

17 Siehe dazu: *Mirsky / Lee*, ACM Computing Surveys 2022, 1.

Generative Adversarial Networks (GAN) oder aber der Personalisierung von Online-Kundenerlebnissen.[18]

Neben den Einsatzgebieten haben sich auch die technischen Zugangsvoraussetzungen für die Erstellung von *Deepfakes* geändert: Waren im Jahr 2017 noch signifikante technische Ressourcen und Expert:innenwissen nötig, um visuell plausible *Deepfakes* zu erzeugen, so ist dies heute mittels vielfältiger, frei verfügbarer Software[19] möglich. Die Nutzung dieser Programme bedarf keines qualifizierten Hintergrundwissens mehr und liefert in kurzer Zeit überzeugende Resultate.[20] Insofern wundert es nicht, dass die Anzahl der im *World Wide Web* kursierenden *Deepfakes* rasant ansteigt.[21]

Vor dem Hintergrund eines zunehmend einfacheren und für Laien zugänglicheren Herstellungsprozesses von KI-generiertem Bild- und Videomaterial einerseits und den stetig wachsenden Einsatzgebieten dieser Technologie andererseits bergen *Deepfakes* – wie alle technischen Entwicklungen – ein bemerkenswertes Potential zur Durchführung krimineller Handlungen. Die Generierung und Nutzung von *Deepfakes* kann strafrechtlich relevant sein, etwa im Kontext von Persönlichkeitsrechtsverletzungen wie Verleumdung und von Delikten wie Erpressung oder Betrug.[22]

3. *Deepfakes in politischen Kontexten*

Im Folgenden wird gezeigt, dass *Deepfakes* und andere Formen von Bild- und Videofälschungen in vielfältigen Varianten auftreten können und dass die Erkennbarkeit von Manipulationen oftmals nicht objektiv messbar ist, sondern auch vom Kontext und der Sensibilität der Betrachter:innen gegenüber Manipulationsrisiken abhängt. *Deepfakes*, die zu Zwecken der Desinformation genutzt werden, bergen das Potential, demokratische Prozesse auf unterschiedliche Art zu beeinflussen**.** Wird die Integrität und

18 *Whittaker* u. a., Australasian Marketing Journal 2021, 204ff.

19 Überblick zu der aktuell gängigen Software: https://beebom.com/best-deepfake-apps-websites/.

20 Vgl. *Riess* in Freiburg (Hrsg.), Täuschungen. Erlanger Universitätstage 2018, 2019, 95.

21 *Ajder u. a.*, The State of Deepfakes, 2019.

22 Siehe für Illustrationen der böswilligen Verwendung von Deepfakes: *Europol, Unicri, Trends Micro,* Report on Malicious Uses and Abuses of Artificial Intelligence (AI), 2020, 52-56.

Fairness demokratischer Wahlen durch den Einsatz von KI-manipulierten oder -generierten Bild- oder Videoinhalten in Frage gestellt, kann dies zu einer Legitimationskrise demokratischer Systeme führen.[23] Für die Wähler:innen birgt die *Deepfake*-basierte Desinformation zu politischen Themen das Risiko, Opfer einer manipulierten Meinungsbildung oder gezielter Verunsicherung zu werden, was sich auch auf ihre Wahlteilnahme und -entscheidung auswirken kann. Für Politiker:innen, deren Auftritte in *Deepfake*-Videos oder -Bildern manipuliert werden, steht ihre Reputation auf dem Spiel, was ihre zukünftigen Wahlchancen beeinträchtigen kann.[24] Insbesondere *Deepfakes*, die darauf abzielen per „microtargeting techniques" bestimmte Personen in Verruf zu bringen, gelingt dieses Unterfangen unter bestimmten Voraussetzungen nachweisbar.[25]

3.1 Einflussnahme auf politische Diskurse durch Deepfakes

Deepfakes reihen sich in das vielfältige technische Repertoire ein, mit dessen Hilfe Meinungen geäußert und Meinungsbildungsprozesse beeinflusst werden können. Im Herbst 2021 verbreiteten sich zahlreiche Variationen eines im Rahmen der Sondierungsgespräche für die Bildung einer Koalitionsregierung auf Bundesebene entstandenen *Selfies* der Führungspersonen von BÜNDNIS 90/DIE GRÜNEN und der FDP.[26] Die kursierenden *Deepfakes* verweisen erkennbar auf das Originalbild. Aufgrund von Inhalt und Aufmachung war für Betrachter:innen unschwer erkennbar, dass diese manipulierten Bilder und Videos nicht echt waren. Der satirische Charakter der unterschiedlichen, durchaus humorvollen Interpretationen des Politiker:innenbildes erschließt sich für durchschnittlich politisch gebildete Betrachter:innen mühelos.

Andere *Deepfakes*, wie das durch digitale Manipulation generierte Video der Sprecherin des Repräsentantenhauses der Vereinigten Staaten Nancy Pelosi,[27] können vom medialen Publikum nicht auf Anhieb als (Ver-)Fälschung identifiziert werden. Der verlangsamte und stockende Sprachfluss der weithin bekannten Demokratin konnte bei den Rezipient:innen den

23 *Krzywoń*, German Law Journal 2021, 676; siehe dazu auch: *Sander*, Chinese Journal of International Law 2019, 1.
24 *Krzywoń*, German Law Journal 2021, 676.
25 *Dobber* u. a., The International Journal of Press/Politics 2020, 69.
26 *Klein*, So lacht das Netz über das FDP-Grünen-Selfie, 2021.
27 *Winkler*, Ein Video zeigt eine betrunkene Nancy Pelosi – und führt uns vor Augen, was mit Deepfakes heute alles möglich ist. 2019.

Eindruck erwecken, die Politikerin stünde unter bewusstseinsverändernden Drogen.[28] Diese Videomanipulation verbreitete sich sehr schnell, so dass sogar die renommierte Nachrichtenagentur Reuters sich im Zugzwang sah, die Manipulation dieses Videos in ihrer Rubrik *Fact Check* auszuweisen.[29]

Eine andere Form der politischen Einflussnahme mittels Manipulation unter (vermeintlicher) Zuhilfenahme von Künstlicher Intelligenz ereignete sich auf der Ebene der russisch-europäischen Beziehungen. Wenige Monate nach der Verhaftung des russischen Oppositionspolitikers Aleksej Naval′nyj im Januar 2021 erreichten mehrere Mitglieder des Europäischen Parlaments Gesprächsanfragen des Naval′nyj-Vertrauten Leonid Volkov.[30] Im Nachgang zu der zustande gekommenen Videokonferenz zwischen den Parlamentsmitgliedern und Volkov kamen Zweifel auf, ob die Person, die als Volkov auftrat, tatsächlich Volkov war.[31] Er selbst erfuhr aus der Presse, dass er am besagten Gespräch teilgenommen haben soll. „Looks like my real face – but how did they manage to put it on the Zoom call? Welcome to the deepfake era …", kommentierte er in den Sozialen Medien den vermeintlich KI-basierten Schwindel mit seiner Identität.[32] Kurze Zeit später bekannte sich das russlandweit bekannte Komiker-Duo Vovan and Lexus, das bereits mehrere Telefongespräche mit hochrangigen Politiker:innen – u.a. gaben Sie sich als das 2019 neu gewählte ukrainische Staatsoberhaupt Volodymyr Zelens′kyj bei einem Telefonat mit dem französischen Präsident Emmanuel Macron aus – erschlichen hatte, zu dem sogenannten *Prank*.[33]

Diese beiden Ereignisse verdeutlichen, dass Politiker:innen stets damit rechnen müssen, dass ihre digitalen Bild- und Videodarstellungen manipuliert werden können. Ob es sich bei einer Darstellungsmanipulation tatsächlich um eine KI-generierte Manipulation handelt, ein technisches

28 *Washington Post*, Faked Pelosi videos, slowed to make her appear drunk, spread across social media, 2019.

29 *Reuters.com*, Fact check: "Drunk" Nancy Pelosi video is manipulated, 2020.

30 *ntv.de u. a.*, In Video-Konferenz getäuscht - Falscher Nawalny-Vertrauter narrt Politiker, 2021; *Roth*, European MPs targeted by deepfake video calls imitating Russian opposition, 2019.

31 *NL Times*, Dutch MPs in video conference with deep fake imitation of Navalny's Chief of Staff, 2021.

32 *Roth*, European MPs targeted by deepfake video calls imitating Russian opposition, 2019.

33 *DerStandard.de*, Russische Scherzbolde legten offenbar Macron mit Telefonstreich rein, 2019.

Mittel anderer Art angewendet oder eine reale Person als „Doppelgänger" eingesetzt wird, erweist sich als zweitrangig.

Deepfakes können aber auch als politisch-künstlerische Intervention inszeniert werden. Im Kontext der US-Präsidentschaftswahlen im Jahr 2020 veröffentlichte die politisch-gesellschaftliche Initiative mit Antikorruptionsfokus *RepresentUs* ein KI-generiertes Video von dem vermeintlich echten nordkoreanischen Staatschef Kim Jong-un. Im Video verweist der ‚Oberste Führer' der Demokratischen Volksrepublik Korea auf die fortgeschrittene Fragilität westlicher demokratischer Strukturen.[34] Die Möglichkeiten, *Deepfakes* in politischen Kontexten zu platzieren und auf diese Weise zu versuchen, Einfluss auf demokratische Prozesse zu nehmen, sind also bereits heute vielfältig.[35] Dieser Trend dürfte mit zunehmenden technischen Möglichkeiten für ausgereifte, schwer erkennbare *Deepfakes* weiter voranschreiten.

3.2 Deepfakes als Form politischer Desinformation

Die Abgrenzung zwischen legitimer kritischer Satire und illegitimer politischer Propaganda wird durch manipulierte oder schlichtweg erfundene Text-, Bild-, Audio- und Videodateien, die im Kontext von *Fake News* verwendet werden, zunehmend erschwert. Tandoc et al. analysierten 34 akademische Beiträge aus den Jahren 2003 bis 2007 und erarbeiteten daraus eine Typologie für *Fake News*. Sie unterscheiden dabei „news satire, news parody, fabrication, manipulation, advertising, and propaganda."[36] Diese Kategorien lassen sich auf die rasant wachsende Anzahl und Vielfalt an *Deepfakes* übertragen. Während die KI-hergestellten Selfie-Variationen aus dem Kontext der deutschen Koalitionsgespräche[37] in den Bereich der „news satire" beziehungsweise „news parody"[38] fallen, ist das Video

34 *RepresentUs*, First Ever Use of Deepfake Technology in a Major Ad Campaign, 2020.
35 In diesem Sinne, siehe zum Beispiel: *Mannheim / Kaplan*, Yale Journal of Law and Technology 2019, 148 ff.
36 *Tandoc Jr. u. a.*, Digital journalism, 2018, 137.
37 *Klein*, So lacht das Netz über das FDP-Grünen-Selfie, 2021.
38 *Tandoc Jr. u. a.*, Digital journalism, 2018, 137.

der vermeintlich betrunkenen Sprecherin des US-amerikanischen Repräsentantenhauses[39] als „manipulation“ [40] zu werten.

Weitere definitorische Arbeit für die Auseinandersetzung mit *Deepfakes* im politischen Kontext leisteten Claire Wardle und Hossein Derakhshan.[41] Systematisch erarbeiteten sie die Bedeutungsgrenzen der Begriffe „mis-information“, „dis-information“ und „mal-information“. Grundsätzlich können *Deepfakes* in jeder dieser Kategorien auftreten. Als inkorrekte Information *ohne* Schädigungsabsicht können sie der Kategorie der „mis-information” zugeordnet werden. Verfolgt die Generierung von *Deepfakes* die Absicht, einer Person, Organisation oder einem Staat zu schaden, indem missverständliche (Teil-)Informationen auf bestimmte Weise miteinander in Verbindung gesetzt werden, können *Deepfakes* als „mal-information” eingestuft werden. Handelt es sich bei *Deepfakes* um „information that is false and deliberately created to harm a person, social group, organization or country“,[42] dann kann eine Bild- und Videomanipulation der Kategorie „dis-information“ zugeordnet werden.

3.3 Die Rolle der Internetnutzer:innen im Kontext desinformierender Deepfakes

Deepfakes reihen sich als neues Phänomen in ein breites Spektrum an Techniken ein, die bereits vor dem Auftreten erster KI-generierter Manipulationen für politische Desinformation genutzt wurden. Die Möglichkeiten, Deepfakes für politische Desinformation zu nutzen, sind indes im Vergleich zu früheren Techniken weitaus größer, wie auch Mannheim und Kaplan betonen: “While ‘Photoshop’ has long been a verb as well as a graphics program, AI takes the deception to a whole new level.”[43]

Die Abgrenzung zwischen „mis-information“, „dis-information“ und „mal-information“ kann im Einzelfall schwierig sein. Das Teilen digitaler Inhalte, deren Authentizität nicht tiefergehend überprüft wurde, ist im digitalen Raum eine gängige Praxis. Dies kann nicht nur urheberrechtliche Fragen aufwerfen, sondern auch dazu führen, dass Internetnutzer:innen

39 *Winkler*, Ein Video zeigt eine betrunkene Nancy Pelosi – und führt uns vor Augen, was mit Deepfakes heute alles möglich ist, 2019.
40 *Tandoc Jr. u. a.*, Digital journalism, 2018, 137.
41 *Wardle / Derakhshan*, Information Disorder: Toward an interdisciplinary framework for research and policy making, 2017.
42 *Wardle / Derakhshan*, Information Disorder: Toward an interdisciplinary framework for research and policy making, 2017.
43 *Mannheim / Kaplan*, Yale Journal of Law and Technology 2019, 148.

unabsichtlich digitale Inhalte verbreiten, die mit einer Schädigungsabsicht hergestellt und im digitalen Raum platziert wurden.

Pennycook et al. haben gezeigt, dass die (angenommene) Echtheit der Informationen bei der Auswahl von Inhalten, die Internetnutzer:innen digital verbreiten, eine nachrangige Rolle spielt.[44] Vorrang bei der Entscheidung für oder gegen das Teilen bestimmter Informationen hat die durch die Veröffentlichung dieser Inhalte antizipierte Aufmerksamkeit für die eigene Internetpräsenz durch andere Internetnutzer:innen.

Der *Code of Conduct on Disinformation*, den die Europäische Kommission 2018 veröffentlicht hat, spricht den Internetnutzer:innen allerdings keine nennenswerte Rolle bei der Verhinderung von Desinformation zu.[45] Das Dokument behandelt hauptsächlich Selbstregulierungsansätze für die Veröffentlichung digitaler Inhalte, denen Privatunternehmen auf freiwilliger Basis folgen können. Ebenso optional formuliert sind die im *Code*[46] enthaltenen Berichtspflichten. Eine gesetzliche Verpflichtung für Unternehmen sieht das Papier nicht vor.

Die Interpretation digitaler Inhalte durch Internetnutzer:innen hängt sowohl vom spezifischen Darstellungskontext der zu beurteilenden Bilder, Videos und sprachlichen Inhalte als auch vom Wissensstand der jeweiligen Nutzer:innen ab. Rössler et al. stellen in diesem Zusammenhang fest, dass Menschen ohne besondere Qualifikation für die Bildevaluierung Fälschungen und Manipulationen in Bildern in 50% der Fälle identifizieren können[47] – statistisch gesehen kommt das Resultat einem zufälligen Raten gleich.[48]

Selbst Fachpublikum lässt sich von KI-manipulierten Bildern in die Irre führen, wie das Szenario um den Beitrag des renommierten norwegischen Fotografen Jonas Bendiksen (Magnum Photos) beim *Visa pour l'image: International Festival of Photojournalism* im Jahr 2021 verdeutlicht. Mittels KI fügte Bendiksen Bären in Bilder einer mazedonischen Industrielandschaft ein. Die Manipulation blieb von der Fachjury unbemerkt.[49] Diese Beispie-

44 *Pennycook u. a.*, Nature 2021, 590.
45 *European Commission*, Code of Practice on Disinformation, 2018.
46 *European Commission*, Code of Practice on Disinformation, 2018.
47 *Rössler* u. a., FaceForensics A Large-scale Video Dataset for Forgery Detection in Human Faces, 2018.
48 *Vaccari / Chadwick,* Social Media + Society, 2020.
49 *Simonite*, A True Story About Bogus Photos of People Making Fake News; *Lyon*, The case for content authenticity in an age of disinformation, deepfakes and NFTs, 2021.

le zeigen, dass Manipulationen in digitalem Bild- und Videomaterial sowohl für Laien als auch Expert:innen schwierig zu erkennen sein können.

4. *Strafbarkeit der Nutzung von Deepfakes im politischen Kontext und Ansätze von Transparenz*

Die Frage nach der Strafbarkeit der Nutzung von *Deepfakes* im politischen Kontext hängt mit der Entscheidung zusammen, ob beziehungsweise unter welchen Umständen die Echtheitsprüfung politischer Aussagen in das Aufgabengebiet von Strafverfolgungsbehörden fallen soll. *Deepfakes* können im Kontext einer politischen Debatte (etwa als Satire) durchaus ein legitimes Ausdrucksmittel sein. Daher stellt sich die Frage, in welchen Fällen demokratische Prozesse dermaßen beeinflusst werden können, dass der Einsatz von Strafrecht als staatliches Kontrollwerkzeug gerechtfertigt wäre. Die strafrechtliche Verfolgung von Deepfakes, die der Kategorie der oppositionellen politischen Meinungsäußerung angehören, wäre problematisch, da die Strafverfolgungsbehörden in vielen Ländern an Weisungen der Regierungen gebunden sind. Grundsätzlich ist die Einflussnahme von Strafverfolgungsbehörden auf diskursive Prozesse in demokratischen Gesellschaften im Hinblick auf die Meinungsfreiheit kritisch zu bewerten. Das Strafrecht sollte hier also nur *ultima ratio* sein.

Vereinzelt reagieren die Gesetzgeber:innen der Welt bereits mit neuen Rechts- und Regulierungsrahmen für politisch desinformierende *Deepfakes*. Im Zusammenhang mit politischen Wahlen erließ Texas als erster US-amerikanischer Bundesstaat ein Gesetz, das politisch motivierte *Deepfakes* in einem klar definierten Zeitraum (30 Tage) vor anstehenden Wahlen verbietet.[50] Auch Frankreich verabschiedete im Jahr 2018 mit dem „Loi relative à la lutte contre la manipulation de l'information"[51] ein Gesetz zur Bekämpfung der Informationsmanipulation. Irreführende Behauptungen und Unterstellungen über politische Akteur:innen und Parteien werden demnach in einem Zeitraum von drei Monaten vor Wahlen unter Strafe gestellt.[52] Soweit *Deepfakes* genutzt werden, um in Wahlkampfzeiten manipulierte und unwahre Informationen zu verbreiten, können sie von diesem Gesetz erfasst sein. Einen ähnlichen Ansatz verfolgt unter

50 *Texas,* Texas Senate Bill 751, 2019.

51 *République Française*, Loi N°2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information.

52 *République Française*, Loi N°2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, Art. L. 163-2, -I.

anderem auch die australische Gesetzgebung mit dem ebenfalls im Jahr 2018 verabschiedeten Gesetz zur Sanktionierung politischer Desinformation, insbesondere im Kontext von Wahlen.[53] Die österreichische Bundesregierung veröffentlichte im Frühjahr 2022 einen „Aktionsplan Deepfake" mit diversen denkbaren Maßnahmen zur Begrenzung der Risiken, die von *Deepfakes* ausgehen.[54]

Für das deutsche Recht vertritt Tobias Lantwin die Auffassung, dass *Deepfakes*, die aus politischen Motiven heraus verwendet werden, unter § 108a StGB (Wählertäuschung) fallen könnten.[55] KI-generiertes Bild- und Videomaterial zeichnet sich jedoch unter anderem dadurch aus, dass es authentisch und integer anmutenden Inhalt mit rein fiktiven Personen oder Ereignissen beinhalten kann. Daher werden *Deepfakes*, deren Inhalt sich *nicht* auf existierende, sondern auf frei erfundene Personen und Geschehnisse stützt, in der Regel nicht unter diesen Straftatbestand fallen. Da Politiker:innen stets auch Privatpersonen sind, besteht sowohl im deutschen[56] als auch im französischen[57] Recht die Möglichkeit, die Herstellung oder Verbreitung von *Deepfake*-Videos wegen der Verletzung von Persönlichkeitsrechten strafrechtlich zu verfolgen, soweit die einschlägigen Straftatbestände erfüllt sind.

Anders stellt sich der Umgang mit *Deepfakes* politischen Inhalts in *nicht*-demokratischen Gesellschaften dar. Autokratische Gesellschaften fokussieren ihren rechtlichen Rahmen nicht auf die Frage, ob *Deepfakes* gegebenenfalls wahre oder unwahre Inhalte vermitteln. Vielmehr steht hier die Konformität beziehungsweise Nonkonformität des *Deepfake*-Inhalts mit der politischen Linie der Regierung im Vordergrund. In diesem Zusammenhang zielt beispielsweise in China ein Gesetzentwurf auf ein Verbot von *Deepfakes* mit nicht regierungskonformem Inhalt ab:

> „Deep synthesis service providers and users shall comply with laws and regulations, respect social mores and ethics, and adhere to the correct

53 National Security Legislation Amendment (Espionage and Foreign Interference) Act 2018, No. 67, 2018; *Douek*, What's in Australia's New Laws on Foreign Interference in Domestic Politics, 2018.

54 *Bundesministerium für Inneres Österreich* (Hrsg.), Aktionsplan Deepfake, 2022.

55 *Lantwin*, MMR 2020, 81.

56 *Ebd.*, 78.

57 Z. B. *République française,* Art. 226-8 Code pénal, 2002. Siehe dazu: *Loiseau*, Légipresse 2020, 64-69.

> political direction, public opinion orientation, and values trends, to promote progress and improvement in deep synthesis services."[58]

Aufgrund des offenen Zugangs zum politischen Diskurs, der demokratische Gesellschaften prägt, sind Demokratien in besonderer Weise für (des-)informationsbasierte Manipulationen anfällig. Zwar muss die Verteidigung demokratischer Grundwerte nicht unweigerlich durch das Mittel des Strafrechts geschehen. Vor dem Hintergrund der zunehmend wachsenden Bedrohungslage durch *Deepfakes*, kann diese Möglichkeit jedoch auch nicht ausgeschlossen werden.[59] Das Spannungsfeld zwischen Meinungs- und Kunstfreiheit einerseits, und der Sicherung einer freien, auf transparentem Informationsfluss basierenden Meinungsbildung andererseits, wird in den kommenden Jahren vor dem Hintergrund der Ausbreitung von *Deepfakes* neu austariert werden müssen. Dabei sollten Instrumente, die Transparenz herstellen und *Deepfakes* als solche erkennbar machen, Vorrang gegenüber strafrechtlichen Sanktionen haben. Diesen Ansatz verfolgt auch die Europäische Kommission in ihrem 2021 vorgelegten Entwurf einer KI-Verordnung, der ein Transparenzgebot für *Deepfakes* als zentralen Regelungsansatz vorschlägt:

> „Nutzer eines KI-Systems, das Bild-, Ton- oder Videoinhalte erzeugt oder manipuliert, die wirklichen Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähneln und einer Person fälschlicherweise als echt oder wahrhaftig erscheinen würden („Deepfake"), müssen offenlegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden."[60]

Der Entwurf schränkt die Transparenzpflicht allerdings für einige Fälle gleich wieder ein: für die Strafverfolgung und für die Nutzung von *Deepfakes* für legitime Zwecke, die von der Meinungs-, Kunst- oder Wissenschaftsfreiheit gedeckt sind.

---

58 chinalawtranslate.com, Provisions on the Administration of Deep Synthesis Internet Information Services (Draft for solicitation of comments), 28. Januar 2022, at https://www.chinalawtranslate.com/en/deep-synthesis-draft/, Art. 4).

59 *Thiel*, ZRP 2021, 202 (205) sieht keinen dringenden Handlungsbedarf; a.A. *Lantwin*, MMR 2019, 578.

60 *Europäische Kommission*, 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, 2021 (Art. 53 Abs. 3).

> „Unterabsatz 1 gilt jedoch nicht, wenn die Verwendung zur Aufdeckung, Verhütung, Ermittlung und Verfolgung von Straftaten gesetzlich zugelassen oder für die Ausübung der durch die Charta der Grundrechte der Europäischen Union garantierten Rechte auf freie Meinungsäußerung und auf Freiheit der Kunst und Wissenschaft erforderlich ist und geeignete Schutzvorkehrungen für die Rechte und Freiheiten Dritter bestehen."[61]

Auch Strategien für die Detektion von *Deepfakes*, wie sie im Verbundprojekt *FAKE-ID* erforscht werden, knüpfen an das Transparenzpostulat an.

5. Projekt FAKE-ID: Interdisziplinäre Erforschung einer Deepfake-Detektion

Auf europäischer und internationaler Ebene werden unterschiedliche Lösungsansätze für den Umgang mit *Deepfakes* verfolgt.[62] In diesem Zusammenhang zielt das interdisziplinäre Forschungsprojekt *FAKE-ID* auf die Erforschung KI-basierter Tools ab, die eine systematische Bewertung der Echtheit von Bild-, Audio- und Videoinhalten technisch unterstützten. Anwendungsfall im Projekt sind gesichtsbasierte Authentifizierungs- und Identifizierungsmethoden.

Formuliert werden zunächst technische Merkmale ‚echter', d. h. nicht manipulierter visueller Medien. Anschließend vergleicht man diese Merkmale mit den Eigenschaften von Bild- und Videobereichen, die mittels Künstlicher Intelligenz verändert oder generiert worden sind. Aufbauend auf diesem Verfahren sieht das Detektionskonzept die Erarbeitung von Kriterien vor, anhand derer KI-manipulierte Bilder und Videodatenströme identifiziert und klassifiziert werden können. Die ermittelten Bild- und Videobereiche, die den Verdacht auf eine Manipulation oder Fälschung nahelegen, erkannte Anomalien und Verdachtsmomente werden anschließend visuell aufbereitet und auf einer Risiko- und Verdachtslandkarte (RVL) dargestellt. Die Markierung der Verdachtsfelder innerhalb von Bildern und Videos soll Anwender:innen in Strafverfolgungsbehörden und

61 Ebd.

62 Dazu gehört auch die Entwicklung eines Deepfakes-Detektors, der eine mögliche Lösung darstellt, siehe z. B. dazu: *Europol's European Cybercrime Centre u. a.*, Report on Malicious Uses and Abuses of Artificial Intelligence (AI), 2020; dazu: *European Parliamentary Research Service*, Tackling deepfakes in European policy, 2021, 24-25.

Gerichten bei der Beurteilung der Authentizität und Integrität von digitalem Bild-und Videomaterial unterstützen.

5.1 Technische und juristische Herausforderungen KI-basierter Deepfake-Detektion

Bei der Konzeption einer *Deepfake*-Detektion stellen sich den Projektteams zahlreiche technische Herausforderungen. Insbesondere gilt es, die Fehlerarten und dazugehörige Fehlerraten der technischen Detektionsmöglichkeiten zu erkennen beziehungsweise die Raten zu optimieren und in den Entscheidungsprozess der menschlichen Anwender:innen miteinzubeziehen. Schließlich stellen die Fehlerarten und -raten der durch die Detektoren produzierten Detektionsfehler höchst relevante Kriterien hinsichtlich der Erklärbarkeit dar, die im Kriterien-Katalog *AIC4 (Artificial Intelligence Cloud Service Compliance Criteria Catalogue*) mit Mindestanforderungen an die sichere Verwendung von Methoden des maschinellen Lernens in Cloud-Diensten festgeschrieben sind.[63] Gefordert wird, dass die Entscheidungen eines Dienstes – im vorliegenden Fall der Detektion von *Deepfakes* – für die Nutzer:innen auf eine Weise dargestellt und kommuniziert werden sollen, die diese Entscheidungen nachvollziehbar macht. Des Weiteren wird festgelegt, dass bei sensiblen Anwendungen (z. B. bei der Nutzung in kritischen Infrastrukturen) die fehlende Erklärbarkeit explizit auszuweisen ist.[64]

Eine weitere technische Hürde bei der Erforschung eines KI-gestützten Detektors stellt der Bedarf an unterschiedlichen Datensätzen dar. Die Trainingsdatensätze, mit denen KI-Systeme ausgearbeitet werden, dürfen nicht dieselben sein, wie diejenigen, die zu Testzwecken verwendet werden. Vielmehr müssen verschiedene real auftretende Charakteristiken einbezogen werden, da ansonsten die Gefahr besteht, ein KI-System zu entwerfen, das nur innerhalb von ‚Laborbedingungen' arbeiten kann. Die fortwährende Notwendigkeit detektierende KI-Systeme anhand aktueller, zunehmend technisch ausgefeilter *Deepfakes* anzupassen, ist dafür prädestiniert, in einem ‚Katz-und-Maus-Spiel' ständiger Qualitätsverbesserung von (a) Deepfakes und (b) Deepfakedetektion zu münden:

63 *Bundesamt für Sicherheit in der Informationstechnik*, Kriterienkatalog für KI-Cloud-Dienste – AIC4, 2021, 29.

64 *Ebd.*, 41.

> "One caution is that the performance of detection algorithms is often measured by benchmarking it on a common data set with known deepfake videos. However, studies into detection evasion show that even simple modifications in deepfake production techniques can already drastically reduce the reliability of a detector" [65]

Aus juristischer Perspektive stellt sich die Frage nach der Rechtskonformität KI-basierter Detektionssysteme. Wenn *Deepfakes* zur Bedrohung demokratischer Prozesse beitragen können, dann birgt ein KI-gestütztes Werkzeug zur *Deepfake*-Erkennung potentiell ebenfalls ernstzunehmende Risiken in Bezug auf die Grundrechte, die Rechtsstaatlichkeit sowie die demokratischen Grundsätze der europäischen Rechtsordnungen.[66] Schließlich unterliegt die Aufgabe der Wahrheitsfindung in erster Linie den Gerichten und nicht den Strafverfolgungsbehörden.

Dieser Problematik wurde in dem 2021 veröffentlichten KI-Verordnungsentwurf der Europäischen Kommission bereits Rechnung getragen. Laut Erwägungsgrund 38 des EU-KI-Verordnungsentwurfs fällt ein KI-System, das auf die Erkennung von *Deepfakes* abzielt, in die Kategorie von KI-Systemen mit hohem Risiko.[67] Eine Studie des Wissenschaftlichen Dienstes des Europäischen Parlaments stuft die Verwendung von KI-basierten *Deepfake*-Detektoren durch die Strafverfolgungsbehörden ebenfalls als hochriskant ein. Diese Klassifizierung basiert darauf, dass die Funktionsweise eines solchen Systems *a priori* nicht ausreichend transparent, erklärbar und dokumentiert ist.[68] Folglich ist damit zu rechnen, dass zukünftig auch die rechtlichen Verpflichtungen verschärft werden, die sich auf detektierende KI-Systeme beziehen. Dies ist auch bei den Forschungen zur *Deepfake*-Detektion im *FAKE-ID*-Projekt zu berücksichtigen.[69]

---

65 *European Parliamentary Research Service*, Tackling deepfakes in European policy, 2021, VIII, S. II-III.

66 *European Commission*, Commission Staff Working Document. Impact Assessment Accompanying the 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, 2021, 49 (unter 5.5, „Impact on the right to freedom of expression").

67 *Europäische Kommission*, 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, 2021, 38.

68 *European Parliamentary Research Service*, Tackling deepfakes in European policy, 2021, 49.

69 *Bundesamt für Sicherheit in der Informationstechnik*, Towards Auditable AI Systems: Current status and future directions, 2021, 21.

5.2 Emanzipatorisches Potential der Deepfake-Detektion für Privatpersonen

Der Wissenschaftliche Dienst des Europäischen Parlaments kommt zu dem Schluss, dass in Zukunft nicht nur staatliche Institutionen, sondern auch Privatpersonen ein ausgeprägtes Maß an Skepsis gegenüber videographischen Informationen entwickeln sollten:

> „[T]he increased likelihood of deepfakes forces society to adopt a higher level of distrust towards all audio-graphic information. Audio-graphic evidence will need to be confronted with higher scepticism and have to meet higher standards. Individuals and institutions will need to develop new skills and procedures to construct a trustworthy image of reality, given that they will inevitably be confronted with deceptive information."[70]

In diesem Sinne erforscht das *FAKE-ID*-Projekt – neben Detektionstools für Strafverfolgungsbehörden und Gerichte – *Deepfake*-Detektionstools für den Gebrauch durch Privatpersonen. Damit könnte der breiten Öffentlichkeit die Möglichkeit geboten werden, KI-generierte Bild- und Videomanipulationen ebenfalls KI-basiert zu identifizieren.

Obgleich die meisten großen sozialen Netzwerke entweder verpflichtet sind[71] oder „sich bemühen",[72] Online-Inhalte, die auf ihren Plattformen verbreitet werden, hinsichtlich einer möglichen Verfälschung zu überprüfen, müssen die Grenzen einer solchen Selbstverpflichtung stets mitbedacht werden. Letztendlich verfolgen Großkonzerne allem voran kommerzielle Ziele, die einer Detektion von *Deepfakes* entgegenstehen können.

6. Fazit

Dieser Beitrag hat gezeigt, dass *Deepfakes* zunehmend ausgereift sind und daher für Betrachter:innen oft nur schwer erkennbar ist, ob Videos und Bilder echt, manipuliert, gefälscht oder sogar frei erfunden sind. Bislang stützen sich die Erkenntnisse über die Risiken, die *Deepfakes* für demokratische Entscheidungsprozesse darstellen können, vorwiegend auf Schil-

---

70 *European Parliamentary Research Service*, Tackling deepfakes in European policy, 2021, VIII.

71 Z. B. *République Française*, Loi N°2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information, Art. L. 163-1.

72 *Facebook Transparency Center*, Kontointegrität und authentische Identität, 2021; *Facebook Transparency Center*, Falschmeldungen, 2021.

derungen von einzelnen Vorkommnissen. Jedoch kann damit gerechnet werden, dass KI-generierte *Deepfakes* und daher auch Manipulationen zunehmend schwer erkennbar sind. Die Herstellung von Transparenz und damit auch die *Deepfake*-Detektion werden infolge dieser Entwicklung zu Instrumenten der Demokratiesicherung.

Trotz der nachvollziehbaren Befürchtungen und Sorgen, insbesondere mit Blick auf demokratische Meinungsbildungsprozesse, die KI in der Gesellschaft hervorrufen, sollten aber auch demokratisierende Potentiale von KI-Anwendungen nicht übersehen werden:

> "Properly designed AI-based accountability tools could probably become the most effective strategy to rebalance the newly structured governance playing field, regain citizens' ownership of democratic decision-making and ensure a community of knowledge and commitment."[73]

Wie Eyal Benvenisti es formuliert, besteht die eigentliche Herausforderung nicht darin, KI als Phänomen unserer Zeit willkommen zu heißen oder abzulehnen. Vielmehr geht es darum, KI-basierte Anwendungen aktiv mitzugestalten. Dabei gilt es, einerseits das technische Potential von KI-gestützten Programmen zu optimieren, andererseits aus einer rechtsstaatlichen Perspektive heraus zu reflektieren, welche Auswirkungen solche KI-basierten Anwendungen auf die Grund- und Menschenrechte sowie auf demokratische Entscheidungsprozesse haben können. Das interdisziplinäre *FAKE-ID*-Projekt verfolgt das Ziel, zur Umsetzung dieses technisch-rechtlich-ethischen Balanceaktes beizutragen.

Grundsätzlich erscheint es möglich, durch KI verursachten Risiken mit ebenfalls KI-basierten Lösungen zu begegnen. Insbesondere in Anbetracht der enormen Geschwindigkeit, mit der riskante KI-Anwendungen entwickelt werden, erscheint es dringend notwendig, KI-basierte Schutzwerkzeuge zu konzipieren. Gleichzeitig gilt es, auch bei der Erforschung und Entwicklung grundrechts- und demokratieschützender KI-Anwendungen die den KI-Tools inhärenten Risiken und Unsicherheiten zu reflektieren und zu minimieren.

---

73 *Benvenisti*, European Journal of International Law 2019, 1089.

Literatur

Ajder, Henry; Patrini, Giorgio; Cavalli, Francesco und Cullen, Laurence (September 2019): The State of Deepfakes: Landscape, Threats, and Impact. Amsterdam: Deeptrace. URL: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf (besucht am 25.02.2022).

Benvenisti, Eyal (2019): Towards Algorithmic Checks and Balances: A Rejoinder. *European Journal of International Law*, 29(4), S. 1087-1090. URL: http://www.ejil.org/archive.php?issue=146 (besucht am 25.02.2022).

Bessette-Symons, Brandy (2018): The robustness of false memory for emotional pictures. *Memory,* 26 (2), S. 171-188.

Bundesministerium für Inneres Österreich (Hrsg.) (2022): Aktionsplan Deepfake. Wien: BMI. URL: https://bmi.gv.at/bmi_documents/2779.pdf (besucht am 14.06.2022).

Bundesamt für Sicherheit in der Informationstechnik (06. Mai 2021): Towards Auditable AI Systems: Current status and future directions. URL: https://www.bsi.bund.de/DE/Service-Navi/Presse/Alle-Meldungen-News/Meldungen/Whitepaper_Pruefbarkeit_KI-Systeme_060521.html (besucht am 25.02.2022).

Bundesamt für Sicherheit in der Informationstechnik (2021): Kriterienkatalog für KI-Cloud-Dienste -- AIC4. URL: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/AIC4/aic4.html (besucht am 25.05.2022).

DerStandard.de (24. Apr. 2019): Russische Scherzbolde legten offenbar Macron mit Telefonstreich rein. URL: https://www.derstandard.de/story/2000101997782/russische-scherzbolde-legten-offenbar-macron-mit-telefonstreich-rein (besucht am 25.02.2022).

Dobber, Tom; Metoui, Nadia; Trilling, Damian; Helberger, Nathalie und de Vreese, Claes (2020): Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?, *The International Journal of Press/Politics* 26(1), S. 69-91, doi:10.1177/1940161220944364.

Douek, Evelyn (11. Juli 2018): What's in Australia's New Laws on Foreign Interference in Domestic Politics. URL: https://www.lawfareblog.com/whats-australias-new-laws-foreign-interference-domestic-politics. (besucht am 25.02.2022).

Duden-Redaktion (2022): Deepfake. URL: https://www.duden.de/rechtschreibung/Deepfake (besucht am 25.02. 2022).

European Commission (26. Sept. 2018): Code of Practice on Disinformation. URL: https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation (besucht am 25.02.2022).

European Commission (21. April 2021): Commission Staff Working Document. Impact Assessment Accompanying the 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, SWD (2021) 84 final. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021SC0084 (besucht am 25.02.2022).

Europäische Kommission (21. Apr. 2021): 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, COM (2021) 206 final.

European Parliamentary Research Service (07. Juli 2021): Tackling deepfakes in European policy, PE 690.039. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf (besucht am 25.02.2022).

Europol's European Cybercrime Centre; United Nations Interregional Crime and Justice Research Institute (UNICRI) und Trend Micro (19. Nov. 2020), Report on Malicious Uses and Abuses of Artificial Intelligence (AI), S. 52-56. URL: https://eucrim.eu/news/report-on-malicious-uses-and-abuses-of-artificial-intelligence/ (besucht am 25.02.2022).

Facebook Transparency Center (zuletzt geändert am 29. Juli 2021): Kontointegrität und authentische Identität. URL: https://transparency.fb.com/de-de/policies/community-standards/account-integrity-and-authentic-identity/ (besucht am 25.02.2022).

Facebook Transparency Center (Stand 01. Okt. 2021): Falschmeldungen. URL: https://transparency.fb.com/de-de/policies/community-standards/false-news/ (besucht am 25.02.2022).

Fährmann, Jan (2018): Drogenpolitik – soziale Kontrolle durch Repressionen? In: Mercer, Milena (Hrsg.), *Altered States*. Berlin: Hatje Cantz Verlag, S. 220-230.

Gerth, Sebastian (2018): Auf der Suche nach Visueller Wahrheit. Authentizitätszuschreibung und das Potenzial der Wirklichkeitsabbildung durch Pressefotografien im Zeitalter digitaler Medien. *IMAGE Zeitschrift für interdisziplinäre Bildwissenschaft*, 14(27), S. 5-23. doi:10.25969/mediarep/16426.

Hardford, Tim (2019): Does pornography still drive the internet? URL: https://www.bbc.com/news/business-48283409 (besucht am 25.02.2022).

Hömberg, Walter und Karasek, Johannes (2008): Der Schweißfleck der Kanzlerkandidatin. Bildmanipulation, Bildfälschung und Bildethik im Zeitalter der digitalen Fotografie. *Communicatio Socialis*, 41(3), S. 276–293.

Hömberg, Walter (2020): Fake News, Medienfälschungen, Grubenhunde. Fälschungsfallen im Journalismus und in den Medien. In: Hohlfeld, Ralf; Harnischmacher, Michael; Heinke, Elfi; Lehner, Lea und Sengl, Michael (Hrsg.): *Fake News und Desinformation: Herausforderungen für die vernetzte Gesellschaft und die empirische Forschung*, Baden-Baden: Nomos, S. 83–96. doi:10.5771/9783748901334.

Isermann, Holger und Knieper Thomas (2010): Bildethik. In: Schicha, Christian und Brosda, Carsten (Hrsg.): *Handbuch Medienethik*, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 304-317.

Klein, Oliver (29. Sept. 2021): Die lustigsten Reaktionen. So lacht das Netz über das FDP-Grünen-Selfie. URL: https://www.zdf.de/nachrichten/panorama/bundestagswahl-vorsondierung-netzreaktionen-100.html (besucht am 25.02.2022).

Krzywoń, Adam (2021): Summary Judicial Proceedings as a Measure for Electoral Disinformation: Defining the European Standard, *German Law Journal* 22(4), S. 673-688. doi:10.1017/glj.2021.23.

Lantwin, Tobias (2019): Deep Fakes – Düstere Zeiten für den Persönlichkeitsschutz. Rechtliche Herausforderungen und Lösungsansätze, *Multimedia und Recht (MMR)*, S. 574-578.

Lantwin, Tobias (2020): Strafrechtliche Bekämpfung missbräuchlicher Deep Fakes: Geltendes Recht und möglicher Regelungsbedarf. *Multimedia und Recht (MMR)*, S. 78-82.

Loiseau, Grégoire (2020): Droits de la personnalité (Janvier 2019 – Décembre 2019). *Légipresse*, 380, S. 64-69.

Lüthe, Rudolf (2007): Die Wirklichkeit der Bilder. Philosophische Überlegungen zur Wahrheit bildlicher Darstellungen. In: Liebert, Wolf-Andreas und Metten, Thomas (Hrsg.): *Mit Bildern lügen*. Köln: Herbert von Halem Verlag, S. 50-64.

Lyon, Santiago (22. Okt. 2021): The case for content authenticity in an age of disinformation, deepfakes and NFTs. URL: https://blog.adobe.com/en/publish/2021/10/22/content-authenticity-in-age-of-disinformation-deepfakes-nfts#gs.mo1gv9 (besucht am 25.02.2022).

Mafi-Gudarzi, Nima (2019): Desinformationen: Herausforderungen für die wehrhafte Demokratie. *Zeitschrift für Rechtspolitik (ZRP)*, S. 65-68.

Mannheim, Karl und Kaplan, Lyri (2019): Artificial Intelligence: Risks to Privacy and Democracy, *Yale Journal of Law and Technology*. 106(21), S. 106-188.

Mirsky, Yisroel und Lee, Wenke (2022): The Creation and Detection of Deepfakes: A Survey, *ACM Computing Surveys*, 54(1), Article No. 7, S. 1-41, doi:10.1145/3425780.

NL Times (24. Apr. 2021): Dutch MPs in video conference with deep fake imitation of Navalny's Chief of Staff. URL: https://nltimes.nl/2021/04/24/dutch-mps-video-conference-deep-fake-imitation-navalnys-chief-staff (besucht am 25.02.2022).

ntv.de; joh und AFP (27. Apr. 2021): In Video-Konferenz getäuscht - Falscher Nawalny-Vertrauter narrt Politiker. URL: https://www.n-tv.de/politik/Falscher-Nawalny-Vertrauter-narrt-Politiker-article22518117.html (besucht am 25.02.2022).

Panholzer, Adrian (28. Juli 2020): Deepfakes wurden durch Pornografie bekannt. URL: https://www.tagesanzeiger.ch/deepfakes-wurden-durch-pornografie-bekannt-635224636195 (besucht am 25.02.2022).

Pennycook, Gordon; Epstein, Ziv; Mosleh, Mohsen; Arechar, Antonio A.; Eckles, Dean und Rand, David G. (2021): Shifting attention to accuracy can reduce misinformation online. *Nature*, 592, S. 590-595. doi:10.1038/s41586-021-03344-2.

Reuters.com (2020): Fact check: "Drunk" Nancy Pelosi video is manipulated. https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI (besucht am 25.02.2022).

RepresentUs (29. Sept. 2020): First Ever Use of Deepfake Technology in a Major Ad Campaign. URL: https://act.represent.us/sign/deepfake-release/ (besucht am 25.02.2022).

Riess, Christian (2019): Die Erzeugung von digitalen Bildfälschungen und ihre Erkennung. In: Freiburg, Rudolf (Hrsg.): *Täuschungen. Erlanger Universitätstage 2018*. Erlangen. FAU University Press, S. 95–114.

Rössler, Andreas; Cozzolino, Davide; Verdoilva, Luisa; Riess, Christian; Thies, Justus und Nießner, Matthias (24. März 2018): FaceForensics A Large-scale Video Dataset for Forgery Detection in Human Faces. URL: https://justusthies.github.io/posts/faceforensics/ (besucht am 25.02.2022).

Roth, Andrew (22. Apr. 2021): European MPs targeted by deepfake video calls imitating Russian opposition. URL: https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition (besucht am 25.02.2022).

Sander, Barrie (2019): Democracy Under The Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections. *Chinese Journal of International Law*, 18(1), 1-56. URL: https://academic.oup.com/chinesejil/article/18/1/1/5359468 (besucht am 25.02.2022).

Simonite, Tom (06. Okt. 2021): A True Story About Bogus Photos of People Making Fake News. URL: https://www.wired.com/story/true-story-bogus-photos-people-fake-news/ (besucht am 25.02.2022).

Steding, Alexander (o.J.): Ein Bild lügt mehr als tausend Worte. URL: https://demokratie.niedersachsen.de/startseite/themen/digitalisierung/fake_news/ein-bild-lugt-mehr-als-tausend-worte-180269.html (besucht am 25.02.2022).

Stefanowitsch, Anatol; Geyken, Alexander; Kopf, Kristin; Kupietz, Marc; Lemnitzer, Lothar und Flach, Susanne (2019): Anglizismus des Jahres 2019. URL: https://www.anglizismusdesjahres.de/anglizismen-des-jahres/anglizismen-des-jahres-adj-2019/ (besucht am 25.02.2022).

Tandoc Jr, Edson C.; Lim, Zheng Wei und Ling, Richard (2018): Defining 'fake news': A typology of scholarly definitions, *Digital journalism,* 6(2,) S. 137-153. doi:10.1080/21670811.2017.1360143.

Thiel, Markus (2021): „Deepfakes" – Sehen heißt glauben? *Zeitschrift für Rechtspolitik (ZRP)*, S. 202-205.

Vaccari, Cristian und Chadwick, Andrew (2020): Deepfakes and Disinformation Exploring the Impact. *Social Media + Society*, 6(1), S. 1-13. doi:10.1177/2056305120903408.

Waddel, Kaveh (07. Januar 2016): How Porn Leads People to Upgrade Their Tech. URL: https://www.theatlantic.com/technology/archive/2016/06/how-porn-leads-people-to-upgrade-their-tech/486032/ (besucht am 25.02.2022).

Wardle, Claire und Derakhshan, Hossein (27. Sept. 2017): Information Disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe report, DGI (2017)09. URL: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c (besucht am 25.02.2022).

Washington Post (2019): Faked Pelosi videos, slowed to make her appear drunk, spread across social media, 2019. https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ (besucht am 25.02.2022).

Webster, Merriam (April 2020): Words We're Watching: 'Deepfake', merriam-webster.com. URL: https://www.merriam-webster.com/words-at-play/deepfake-slang-definition-examples (besucht am 25.02.2022).

Weiffen, Nicole (2019): Der Anglizismus des Jahres 2019 lautet "... for future", URL: https://www.duden.de/presse/anglizismus-des-jahres-2019 (besucht am 25.02.2022).

Whittaker, Lucas; Letheren, Kate und Mulcahy, Rory (2021): The Rise of Deepfakes: A Conceptual Framework and Research: Agenda for Marketing. *Australasian Marketing Journal* 29(3), S. 204-2014, doi:10.1177/1839334921999479.

Winkler, Peter (25. Mai 2019): Ein Video zeigt eine betrunkene Nancy Pelosi – und führt uns vor Augen, was mit Deepfakes heute alles möglich ist. URL: https://www.nzz.ch/international/deep-fakes-nancy-pelosi-video-manipuliert-ld.1484614 (besucht am 25.02.2022).