Aktionsplan Deepfake

Aktionsplan Deepfake

Wien, 2022

Bundeskanzleramt

Bundesministerin für
Frauen, Familie, Integration und Medien

Bundesministerium
Europäische und internationale
Angelegenheiten

Bundesministerium
Justiz

Bundesministerium
Landesverteidigung

Bundesministerium
Inneres

Impressum

MedieninhaberIn, VerlegerIn und HerausgeberIn:
Bundesministerium für Inneres
Herrengasse 7, 1010 Wien
bmi.gv.at
AutorInnen: Abteilung I/11
Layout: Abteilung I/5/b
Druck: Digitalprintcenter des
Bundesministeriums für Inneres
Herrengasse 7, 1010 Wien
Wien, 2022

Präambel

Dieser Aktionsplan wurde im Rahmen einer interministeriellen Arbeitsgruppe bestehend aus Vertreterinnen und Vertretern des Bundeskanzleramts (BKA), des Bundesministeriums für europäische und internationale Angelegenheiten (BMEIA), des Bundesministeriums für Justiz (BMJ), des Bundesministeriums für Landesverteidigung (BMLV) unter Federführung des Bundesministeriums für Inneres (BMI) erstellt. Der Aktionsplan soll dem Ministerrat vorgestellt werden und in Folge dem Nationalrat vorgelegt werden.

Inhalt

1 Auftrag

1.1 Parlamentarischer Auftrag

Das **Parlament** hat den Entschließungsantrag Nr 365/A(E) vom 27. Februar 2020 am 14. Oktober 2020 einstimmig angenommen. Der Antrag fordert:

„Die Bundesregierung, insbesondere der Bundesminister für Inneres, wird aufgefordert, dem Nationalrat ein Konzept zum Umgang mit Deepfakes und einer Strategie zur Bekämpfung von politischen, gesellschaftlichen und wirtschaftlichen Risiken durch Deepfakes vorzulegen. Darin soll das Ziel verfolgt werden, frühzeitig Maßnahmen zur Eindämmung von Risiken zu finden."

Österreich wird durch einen gesamtstaatlichen Ansatz der zuständigen Bundesministerien Sicherheit und Resilienz gegen Gefährdungen durch Deepfakes herstellen.

1.2 Auftrag im Regierungsprogramm 2020-2024

Das **Regierungsprogramm 2020-2024** setzt sich an mehreren Stellen mit dem Thema Desinformation und digitale Kriminalitätsbekämpfung auseinander:

- Kapitel Medien: „Schutz vor Desinformation"
- Kapitel Europa: „Verstärkter Kampf gegen Desinformation und Wahlbeeinflussung auf allen Ebenen. Stärkere Zusammenarbeit bei Cybersicherheit von allen betroffenen Ressorts der Bundesregierung und bestehende Mechanismen der EU wie Frühwarnsystem und Taskforce zur Früherkennung von Desinformationskampagnen stärken und mehr nutzen."
- Kapitel Europa: „Entwicklung einer neuen EU-Digitalstrategie mit gemeinsamen Schwerpunkten, in denen Europa künftig den globalen Fortschritt anführen und von anderen Akteurinnen und Akteuren unabhängig werden soll, z. B. künstliche Intelligenz, Internet der Dinge, Cybersicherheit"
- Kapitel Justiz: „Stärkung von Sicherheit, Rechtsfrieden und des Schutzes der höchsten Rechtsgüter, nicht nur in der analogen Welt, sondern auch in der digitalen Welt", „Bündelung staatsanwaltlicher Ermittlungskompetenzen zur Bekämpfung digitaler Verbrechen"
- Kapitel Justiz: „Einsetzung einer ressortübergreifenden Taskforce zur effizienten Bekämpfung von Hass im Netz und anderer digitaler Kriminalitätsformen"

1.3 Erweiterte Handlungsgrundlage

- EU-Digital Services Act vom 15. Dezember 2020
- EU-Aktionsplan für Demokratie vom 3. Dezember 2020
- EU-The Landscape of Hybrid Threats: A Conceptual Model vom 30. November 2020
- Empfehlung des Nationalen Sicherheitsrates vom 11. September 2019 zur Erhöhung der Kapazitäten im Bereich Screening und Monitoring möglicher schädlicher Aktivitäten
- EU-Aktionsplan gegen Desinformation vom 5. Dezember 2018
- EU-Gemeinsamer Rahmen für die Abwehr hybrider Bedrohungen vom 6. April 2016

2 Definition

Der Begriff „Deepfake“ wird als Überbegriff für verschiedene Formen der audiovisuellen Manipulation einschließlich Video, Audio oder beides verwendet. Typischerweise wird zur Erstellung von Deepfakes eine auf Künstlicher Intelligenz (KI) basierte Technologie verwendet. Deepfakes sind perfekt gefälschte Videos, Bilder oder Audio in denen Personen Aussagen in den Mund gelegt werden oder in denen sie scheinbar Handlungen begehen, die in Wirklichkeit nie stattgefunden haben. Mit leistungsfähigen Verfahren der KI lassen sich Videos in einer Weise manipulieren, sodass zumindest mit bloßem Auge nicht mehr zu erkennen ist, ob sie echt sind oder manipuliert wurden.[1] Der Begriff Deepfake setzt sich aus den Begriffen „deep learning“ und „fake“ zusammen. Deep learning ist eine spezielle KI-Technik und „fake“ steht für Fälschung oder Falschmeldung. Deepfakes fügen sich ein in die lange Reihe der medialen Manipulationen zum Zweck der Falsch- oder Desinformation. In Folge wird eine Abgrenzung zu diesen Bereichen vorgenommen.

2.1 Erläuterung zu Fake News/Deepfake

Fake News sind in den Medien und im Internet, besonders in sozialen Netzwerken, in manipulativer Absicht verbreitete Falschmeldungen. In Abgrenzung zur Desinformation geht es hier um jede Nachricht, auch solche, die ohne Absicht, ein gewisses Ziel zu erreichen, verbreitet werden (siehe unten Desinformation). Deepfakes fallen unter den Begriff der Fake News, wenn sie mit der Absicht hergestellt werden, zu manipulieren. Schon der Begriff Deepfake lässt darauf schließen, dass diese Form der Manipulation unter den Begriff der Fake News fällt.

2.2 Erläuterung zu Desinformation/Deepfake

Unter Desinformation versteht man eine falsche und zielgerichtet erzeugte irreführende Information, die verbreitet wird, um ein gewisses politisches oder wirtschaftliches Ziel zu erreichen oder einer Person, sozialen Gruppe, Organisation oder Land zu schaden. Die Absicht der Erreichung des Ziels bzw. des Schadens ist hier das ausschlaggebende Kriterium. Sollten Deepfakes mit der Absicht hergestellt werden einer präzisen Gruppe Schaden hinzuzufügen, indem dieser Gruppe manipulative falsche Nachrichten in Form von Deepfakes zugeführt werden, dann fallen sie in den Bereich der Desinformation (Deepfakes als Werkzeug der Desinformation). Ein durch Desinformation mitgeprägtes Meinungsklima stellt letztendlich eine Bedrohung der Gesellschaft und der Demokratie dar. Wenn es beispielsweise um ein rein satirisches Deepfake geht, fällt es nicht in den Bereich der Desinformation.

[1] Norbert Lossau, „Deepfake: Gefahren, Herausforderungen und Lösungswege“, KAS, Februar 2020, https://www.kas.de/de/analysen-und-argumente/detail/-/content/deep-fake-gefahren-herausforderungen-und-loesungswege.

Beispiel

Veröffentlichung eines **satirischen Deepfake-Videos** einer Politikerin oder eines Politikers oder eines/einer Vorsitzenden einer Religionsgemeinschaft.

Antwort

- Bei der Regulierung von Deepfake-Videos sind die relevanten Grund- und Persönlichkeitsrechte zu berücksichtigen und ist insbesondere auf den besonderen Schutz der Meinungsäußerungsfreiheit und der Kunstfreiheit zu achten
- Auch hier wird, wie beim zivilrechtlichen Persönlichkeitsschutz im Allgemeinen, eine Interessenabwägung bzw. eine Prüfung auf der Rechtfertigungsebene vorzunehmen sein. Mögliche Rechtfertigungsgründe wären etwa die Einwilligung des Betroffenen, die Meinungsäußerungsfreit sowie andere Grundrechte oder Persönlichkeitsrechte. Im Rahmen der Interessenabwägung wäre insbesondere die Kunstfreiheit (Art. 17a StGG, Art. 13 GRC) zu berücksichtigen, die auch satirische Darstellungen schützt. Dabei ist zu beachten, dass Politikern sowie Personen des öffentlichen Lebens (sog „public figures“) von Rechtssprechung und Literatur im Allgemeinen nur ein eingeschränkter Persönlichkeitsschutz zugestanden wird, je nachdem ob die Veröffentlichung zu einer Debatte von allgemeinem gesellschaftlichen Interesse beiträgt oder bloß zur Befriedigung der Neugier eines bestimmten Publikums dient. Grenzen finden sich beim Wertungsexzess, bei Vorwürfen ohne Tatsachensubstrat und bei Berührung des höchstpersönlichen Lebensbereichs (vgl. Meissel in Klang^{3} § 16 ABGB Rz 102)
- Legitimer Einsatz der Technologie in Bereichen wie Wissenschaft, Kunst, Bildung, Medien
- Kennzeichnung von Deepfakes im Bereich der Kunst- und Meinungsäußerungsfreiheit sollten ermöglichen, dass es zu keinerlei Einschränkung dieser Kunstform bzw. der Grundfreiheit kommt. Es geht der Satire grundsätzlich nicht darum zu täuschen, im Gegensatz zur Desinformation

2.3 Erläuterung zu Hybride Bedrohungen/Deepfakes

Hybride Bedrohungen kombinieren konventionelle und unkonventionelle, militärische und nichtmilitärische Aktivitäten, die von staatlichen oder nichtstaatlichen Akteurinnen und Akteuren koordiniert eingesetzt werden, um bestimmte politische Ziele zu erreichen. Die Bandbreite der hybriden Bedrohungen reicht von Cyberangriffen auf öffentliche und wirtschaftliche Ziele, gezielten Desinformationskampagnen, Störungen kritischer Infrastrukturen bis hin zu feindlichen militärischen Aktionen. Deepfakes sind in diesem Zusammenhang nur eine Aktivität. Sie müsste mit einer anderen oben genannten Aktivität kombiniert werden, um zu einer hybriden Bedrohung zu werden.

3 Bedrohungslage

Einem Bericht der Firma Deeptrace zufolge waren im Jahr 2019 14.678 Deepfake-Videos online. Dabei war die Verteilung der Videos zwischen pornographischen und nicht pornographischen Inhalts 96 zu 4 Prozent.[2] Man muss davon ausgehen, dass es zum jetzigen Zeitpunkt schon viel mehr solcher Videos gibt.[3] Cyberkriminalität hat in den letzten Monaten sprunghaft zugenommen, nicht zuletzt bedingt durch die Corona-Pandemie. Dieser Trend wird sich weiter verschärfen.[4] In Folge wird die Bedrohungslage an Hand von (nicht taxativen) Szenarien beschrieben, die sich in drei Teilbereiche aufteilen: 1. Sicherheitspolitik, 2. Recht, 3. Technik.

3.1 Sicherheitspolitik

Szenario 1

Ein falsches Video mittels KI hergestellt, um mit politischer Absicht zu täuschen oder zu manipulieren, kann eine erhebliche Gefahr für die Integrität einer Demokratie darstellen. Als Beispiel sei ein Video genannt, das ein Staatsoberhaupt oder ein Regierungsmitglied zeigt, das Dinge sagt, die dann in Folge zu Massendemonstrationen sowie **Regierungs- und Staatskrise** führen. Wenn sich diese Elemente im Internet verbreiten, stören sie demokratische Prozesse und das Vertrauen der Bürgerinnen und Bürger.

Antwort

- Effektiven und koordinierten Krisenmanagementmechanismus aufbauen
- Sensibilisierung der Bevölkerung
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfakes

Szenario 2

Ein einziges spektakuläres Video, das ausgesprochen realistisch ist, führt zu einer Kette von Reaktionen in anderen Bereichen wie dem **Zusammenbruch der Aktienmärkte oder Wahlbeeinflussung**.[5]

---

[2] Deeptrace, “The State of Deepfakes”, Landscape Threats and Impact”, September 2019, S 1.

[3] Will Douglas Heaven, “Facebook just released a database of 100,000 deepfakes to teach AI how to spot them”, 2. Juni 2020, MIT Technology Review, https://www.technologyreview.com/2020/06/12/1003475/facebooks-deepfake-detection-challenge-neural-network-ai/.

[4] Security Insider, 12.01.2021, „Mehr Desinformation mit Deepfakes und mehr Corona-Betrug“, https://www.security-insider.de/mehr-desinformation-mit-Deepfakes-und-mehr-corona-betrug-a-990717/.

[5] Alex Engler, “Fighting Deepfakes when detection fails”, Brookings, 14. November 2019, https://www.brookings.edu/research/fighting-Deepfakes-when-detection-fails/.

Antwort

- Effektiven und koordinierten Krisenmanagementmechanismus aufbauen
- Sensibilisierung der Wirtschaft über Folgen von Deepfakes
- PPP Modelle
- Sensibilisierung der Bevölkerung
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfakes
- Siehe Pkt. 3.2. Recht

Szenario 3

Desinformationskampagnen durch ausländische Akteurinnen und Akteure: Man muss davon ausgehen, dass die Technologie sowohl von anderen Staaten als auch von terroristischen Organisationen für ihre Zwecke eingesetzt werden wird. Desinformation nimmt hier eine neue Dimension ein. Das Ziel von Desinformationskampagnen ist die Erreichung einer Informationsüberlegenheit unterhalb der Schwelle eines bewaffneten Angriffes.

Antwort

- Internationale Zusammenarbeit
- Zusammenarbeit auf EU-Ebene
- Nationale Umsetzung des European Democracy Action Plan
- Enge Einbindung von Medien und Providern von sozialen Netzwerken zur Schaffung eines gesamtgesellschaftlichen Zugangs
- Einbeziehung von Vertretern der Zivilgesellschaft einschließlich NGOs

Szenario 4

Die Gesellschaft und insbesondere die Medien stehen vor neuen Herausforderungen. Es kann zu einer kontinuierlichen **Erosion des Vertrauens** in digitale Inhalte kommen. Wenn eine große Menge an Deepfake-Videos (politischen Inhalts) von Amateurinnen und Amateuren hochgeladen wird, deren Verbreitung nicht mehr eingedämmt werden kann, dann kann dies zur Infragestellung staatlicher Institutionen führen oder auch zur Beeinflussung von Wahlen. Faktoren, die die Bedrohung beschleunigen sind Videoplattformen oder private Messaging-Plattformen. Nach dem gleichen Prinzip wie Fake News spielt die

Authentizität eines Videos oft keine Rolle mehr, nachdem es mehrfach über Social Media geteilt wurde.

Antwort

- Sensibilisierung der Bevölkerung
- Aktiver Austausch mit Medien und unabhängigen Faktenprüfern
- Koordinierter Austausch mit Internetplattformen, Providern und unabhängigen Fact-Checker Plattformen

Szenario 5

Zur konkreten Bedrohung für die Innere Sicherheit können gefälschte Videos werden, die bspw. **Polizeigewalt** unterstellen oder vermeintlich Polizistinnen und Polizisten präsentieren, die strafbare Handlungen begehen. Diese können zum Verlust des Vertrauens in die Polizei und den gesamten Rechtsstaat führen.

Antwort

- Effektiven und koordinierten Krisenmanagementmechanismus aufbauen
- Sensibilisierung der Bevölkerung u.a. durch Bildungseinrichtungen
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfakes
- Siehe Pkt. 3.2. Recht

3.2 Recht

Vorauszuschicken ist, dass der Versuch, dem Phänomen des Deepfake im Internet mit rechtlichen Instrumenten beizukommen, an den wirklichen Problemen vorbeigehen könnte. In aller Regel wird es zwar ohnedies Rechtsgrundlagen für zivilrechtliche/strafrechtliche Ansprüche gegen die Herstellung und Verbreitung von Deepfakes geben, diese werden aber angesichts einer massenhaften, anonymen und nicht als Deepfakes offengelegten Verbreitung im Internet praktisch schwer durchsetzbar sein. Es wird daher eine staatliche Intervention auf international möglichst breiter Basis empfohlen (Europa, Vereinte Nationen), um dem Phänomen mit geeigneten Maßnahmen zu begegnen. Zentrale Frage sollte daher sein, ob ein auf internationaler Ebene akkordiertes, mit Verwaltungsstrafen sanktioniertes Verbot der zur Herstellung von Deepfakes eingesetzten Technologien angesichts der damit verbundenen Gefahren gerechtfertigt und nötig ist.

Eine engere Kooperation auf europäischer Ebene in der Deepfake Problematik wäre empfehlenswert. Das Rapid Alert System der EU könnte zu einem raschen Austausch genützt werden.

Hinsichtlich materiellrechtlicher und prozessualer Aspekte des Strafrechts wird aktuell kein Regelungs- bzw. Handlungsbedarf gesehen; der Rechtsbestand ist für entsprechende Szenarien ausreichend ausgestaltet, zumal Deepfake-Videos geeignet sind, alle möglichen Delikte zu erfüllen, etwa Betrug oder gefährliche Drohung oder Erpressung, indem das Video zur Täuschung oder Drohung eingesetzt wird. Dadurch ist der Täter unter Verwendung solcher Videos herkömmlich wegen Betruges, gefährlicher Drohung oder Erpressung, etc. strafbar.

Bei der Regulierung von Deepfake-Videos sind die relevanten Grund- und Persönlichkeitsrechte zu berücksichtigen und ist insbesondere auf den besonderen Schutz der Meinungsäußerungsfreiheit und der Kunstfreiheit zu achten.

Szenario 1

Die Möglichkeiten für Kriminelle erreichen eine neue Dimension. ENISA geht in ihrem Cyberbedrohungsbericht 2020 davon aus, dass **Cyberkriminelle** verstärkt zu Deepfakes greifen werden, um Unternehmen zu erpressen.[6] Der sog CEO Fraud oder Erpressungen im Allgemeinen werden mithilfe von Deepfake-Technologien eine neue Dimension annehmen. Laut Bundeskriminalamt ist dies in der Praxis in ersten Ansätzen bereits zu beobachten.

[6] ENISA, Threat Landscape 2020, 20.Oktober 2020, https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends, S 15.

Antwort

- §§ 105f. StGB (schwere) Nötigung, § 107 StGB gefährliche Drohung, § 144 StGB Erpressung, §§ 146ff. StGB Betrug, § 148a StGB Betrügerischer Datenmissbrauch
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 2

Verwenden des Gesichts einer bestimmten (unbeteiligten) Person für die Produktion eines **Fake-Pornofilms**.[7] Betroffen sind vor allem Frauen und Kinder.[8]

Antwort

- § 107c StGB Fortdauernde Belästigung im Wege einer Telekommunikation oder eines Computersystems („Cybermobbing“)
- Schutz des **Rechts am eigenen Bild**. Die Bestimmung über den Bildnisschutz, der in § 78 UrhG normiert ist, stellt dabei auf die „berechtigten Interessen“ des Abgebildeten ab, und soll vor allem davor schützen, dass jemand durch die Verbreitung seines Bildnisses bloßgestellt, sein Privatleben der Öffentlichkeit preisgegeben wird oder sein Bildnis auf eine Art benützt wird, die zu Missdeutungen Anlass geben kann oder entwürdigend oder herabsetzend wird (vgl Aicher in Rummel/Lukas, ABGB 4 § 16 Rz 24). Eine Verletzung dieses Persönlichkeitsrechts ist bei Verbreitung eines Deepfakes mit pornographischem Inhalt (gegen den Willen der abgebildeten Person) ohne Zweifel gegeben
- Ev. auch betroffen: **Recht an der eigenen Stimme**, das **Recht am gesprochenen Wort**, das **Recht auf Privatsphäre** sowie das **Recht auf geschlechtliche Selbstbestimmung**. Interessenabwägung bzw. eine Prüfung auf der Rechtfertigungsebene notwendig
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 3

Verwenden des Gesichts einer bestimmten (unbeteiligten) Person für die Produktion eines Deepfake-Videos, bspw. ein Video, das die Person beim Begehen einer **strafbaren Handlung** zeigt, um die konkrete Person zu diskreditieren oder zu erpressen. Lt. Bundeskriminalamt ist dies in der Praxis in ersten Ansätzen bereits zu beobachten.

[7] Bis dato laut Bundeskriminalamt Fälle, bei denen Kindergesichter mit Photoshop in inkriminierte Fotos eingefügt wurden. Von der Machart nicht so professionell, waren die Manipulationen in diesen Fällen als solche erkennbar. Wenn die Software das hält, was manche versprechen und Fälschungen mit freiem Auge nicht mehr zu erkennen sind, ist in der Praxis mit einem nicht außeracht zu lassenden Problem zu rechnen.

[8] Deeptrace, „The State of Deepfakes – Lanscapes, Threats and Impact“, September 2019.

Antwort

- § 144 StGB Erpressung, § 107c StGB Cybermobbing, §§ 297 StGB Verleumdung
- Schutz des Rechts am eigenen Bild (siehe Szenario 2)
- Eventuell auch betroffen: **Recht an der eigenen Stimme**, **Recht am gesprochenen Wort**, **Recht auf Privatsphäre**. Interessenabwägung bzw. eine Prüfung auf der Rechtfertigungsebene notwendig
- Ehrenbeleidung und Kreditschädigung nach § 1330 ABGB
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 4

Bedrohung mittels Veröffentlichung eines Videos über ein Unternehmen, damit bspw. dessen Aktienkurse abstürzen.

Antwort

- siehe Szenario 1
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 5

Kriminelle setzen die Technologie ein, um **Überwachungsvideos** zu manipulieren. Sobald sich Videos mittels KI-Verfahren in einer Weise verändern lassen, dass Manipulationen zumindest mit bloßem Auge nicht mehr zu erkennen sind, stehen die Strafverfolgungsbehörden vor einem „Beweiskraft-Problem". Oftmals sind Videoaufnahmen von Verdächtigen (z.B. bei Einbrüchen in Bankschließfächer) wichtige - wenn nicht sogar die wichtigsten - Ermittlungsansätze. Gelingt es den Tätern die Überwachungsvideos entsprechend zu manipulieren, werden die Ermittlungen dadurch extrem erschwert bzw. laufen unter Umständen sogar ins Leere.

Antwort

- § 126a StGB Datenbeschädigung, § 225a StGB Datenfälschung, ggf. § 293 StGB Beweismittelfälschung
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake
- Sensibilisierung, ggfs Schulung der ermittelnden Beamtinnen und Beamten, Staatsanwältinnen und Staatsanwälte, Richterinnen und Richter

Szenario 6

Veröffentlichung eines Deepfake-Videos, um eine **Wahl** oder Volksabstimmung gezielt zu beeinflussen.

Antwort

- § 263 StGB Täuschung bei einer Wahl oder Volksabstimmung, § 264 StGB Verbreitung falscher Nachrichten bei einer Wahl oder Volksabstimmung
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 7

Gefälschtes Bildmaterial kann als **digitaler Beweis** für beliebige Situationen genutzt werden

Antwort

- Sensibilisierung, ggfs Schulung der ermittelnden Beamtinnen und Beamten, Staatsanwältinnen und Staatsanwälte, Richterinnen und Richter
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 8

Eine verdächtige Person hat auf ihrem Mobiltelefon ein Video, welches sie bei einer bekannten Veranstaltung in z.B. Bregenz zeigt, während sie in Wien eine Straftat begeht. Es besteht die Gefahr, dass mangels Wissens der Ermittlerinnen und Ermittler, Staatsanwältinnen und Staatsanwälte um die Möglichkeiten des Deepfakes die **Echtheit des Videos nicht hinterfragt** wird und dieses damit in ungerechtfertigter Weise zur Entlastung der beschuldigten Person beiträgt. Deepfakes könnten auch gezielt eingesetzt werden, um den Verdacht auf eine in Wirklichkeit unbeteiligte Person zu lenken.

Antwort

- Sensibilisierung, ggfs Schulung der ermittelnden Beamtinnen und Beamten, Staatsanwältinnen und Staatsanwälte, Richterinnen und Richter
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 9

„**Deep Nude**“: Im Jahr 2019 wurde eine App geschaffen, die es ermöglichte, eine Frau auf einem Foto zu entkleiden und nackt zu zeigen. Die App ging rasch wieder offline, Nachahmung ist durchaus möglich.

Antwort

- § 107c StGB „Cybermobbing“
- Schutz des Rechts am eigenen Bild (siehe Szenario 2)
- Eventuell auch betroffen: das **Recht auf Privatsphäre** sowie das **Recht auf geschlechtliche Selbstbestimmung**. Interessenabwägung bzw. eine Prüfung auf der Rechtfertigungsebene notwendig
- Sensibilisierung, ggfs Schulung der ermittelnden Beamtinnen und Beamten, Staatsanwältinnen und Staatsanwälte, Richterinnen und Richter
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 10

Spionage[9]: 2019 wurde ein LinkedIn-Konto aufgedeckt, das sich als „Katie Jones“ ausgab, Forscherin eines führenden US-Think Tanks. Sie (künstlich erzeugte Person) war jedoch Teil einer Spionageoperation. Zum Zeitpunkt der Entfernung durch LinkedIn im Juni 2019 hatte die Person bereits mit mehreren Mitarbeiterinnen und Mitarbeitern der Regierung Kontakt.

Antwort

- Sechzehnter Abschnitt des StGB (§§ 252ff StGB)
- Sensibilisierung, ggfs Schulung der ermittelnden Beamtinnen und Beamten, Staatsanwältinnen und Staatsanwälte, Richterinnen und Richter
- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake

Szenario 11

Wenn im Rahmen bspw. einer Terrorlage Deepfakes in Umlauf gelangen, können diese die Rekonstruktion eines Tatherganges erschweren und die Identifikation/Festnahme eines unter Umständen noch flüchtigen Täters oder einer Täterin erschweren.

[9] Satter Raphael, “Experts: Spy used AI-generated face to connect with targets”, AP News, https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d.

Antwort

- Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfake
- Sensibilisierung der Bevölkerung
- Zusammenarbeit mit Social Media Plattformen

Nicht zuletzt könnte die Herstellung und Verbreitung eines Deepfakes auch im Wege des neuen, durch das **Hass-im-Netz-Bekämpfungs-Gesetz** mit 1. Jänner 2021 eingeführten Mandatsverfahrens nach § 549 ZPO bekämpft werden, wenn es sich um eine „erhebliche, eine natürliche Person in ihrer Menschenwürde beeinträchtigenden Verletzung von Persönlichkeitsrechten" handelt. Bei der Verletzung der Menschenwürde geht es um solche Eingriffe, für die prima vista eine Rechtfertigung nicht vorliegt. Die Menschenwürde einer Person wird verletzt, wenn diese entmenschlicht (das Recht auf Menschsein schlechthin abgesprochen wird) bzw. herabgewürdigt oder erniedrigt wird. Praktische Anwendungsfälle wären Hasspostings, in denen Todes- oder Vergewaltigungswünsche ausgesprochen werden, aber auch, wie in den Materialien zum Hass-im-Netz-Bekämpfungs-Gesetz angesprochen (ErläutRV 481 BlgNR 27. GP 12), das heimliche Anfertigen kompromittierender Bildaufnahmen („Upskirting" § 120a StGB). Bei einer unfreiwilligen Manipulation zum Zweck der Pornographie wird eine solche Verletzung der Menschenwürde in aller Regel anzunehmen sein.

Das **Urheberrecht** könnte etwa dann angesprochen sein, wenn ein bereits vorhandenes Video verwendet wird. Geltendes **Datenschutzrecht** bzw. die DSGVO sind anwendbar, weil die Voraussetzungen rechtmäßiger Datenverarbeitungen auch für gefälschte Daten erfüllt sein müssen.

Weiters wird darauf hingewiesen, dass der Vorschlag eines **Digital Services Act** der Europäischen Kommission in Art. 26 eine Verpflichtung für sehr große Online-Plattformen enthält, eine Risikobewertung vorzunehmen, die auch die vorsätzliche Manipulation ihrer Dienste umfasst, die negative Auswirkungen auf den Schutz der öffentlichen Gesundheit, Minderjähriger, Diskurs der Zivilgesellschaft, oder auf Wahlen und die öffentliche Sicherheit hat oder haben könnten. Bei solchen Risiken sind nach Art. 27 Maßnahmen zu ergreifen. Diese Maßnahmen könnten eine zusätzliche Antwort zu einigen der aufgezeigten Szenarien liefern.

Ein detailliertes Einbeziehen ethischer, demokratiepolitischer sowie **grund- und menschenrechtlicher** Aspekte (einschließlich Kinder-und Jugendschutz) sowie entsprechende internationale Zusammenarbeit auf EU- und internationaler Ebene ist im Bereich der Deepfake Problematik sehr wichtig.

3.3 Technik

Szenario 1

Zwischen dem Herstellen und Verbreiten von Deepfakes und dem Erkennen und Gegensteuern herrscht, analog wie auf dem Gebiet der klassischen IT-Security, ein permanenter Wettlauf, bei dem die angreifende Person oft einen kleinen Schritt voraus ist. Vor allem geben auch die Weiterentwicklungen der KI-Technologie den Erstellern von Deepfakes neue Werkzeuge in die Hand, um immer schwerer erkennbare Fälschungen zu produzieren.

Antwort

- Regelmäßige und gezielte Weiterentwicklung der eingesetzten Werkzeuge zur Erkennung von Deepfakes, um präventive und reaktive Maßnahmen ergreifen zu können
- Zusammenarbeit mit Forschung und Entwicklung
- Internationale Zusammenarbeit

Szenario 2

Künstliche Persönlichkeiten (**Artificial Personas**) sind eine große Bedrohung, die durch das Zusammenwachsen diverser KI-Technologien entstehen, vor allem durch die Kombination von qualitätsmäßig verbesserter KI-gesteuerter Texterzeugung und Chatbots. Im Medienbereich gibt es bereits erste Anwendungen, die für die Bereiche Sport und Finanzen (Börsenberichte) Nachrichten maschinell erzeugen. Auch im E-Commerce Bereich werden Chatbots immer öfter eingesetzt.

Antwort

- Regelmäßige und gezielte Weiterentwicklung der eingesetzten Werkzeuge zur Erkennung von Deepfakes, um präventive und reaktive Maßnahmen ergreifen zu können
- Zusammenarbeit mit Forschung und Entwicklung
- Rechtliche Abklärung
- Internationale Zusammenarbeit

Szenario 3

Der Einsatz und die Weiterentwicklung von Generative Adversarial Networks (GANs) wird insbesondere im Bereich Deepfakes beobachtet werden müssen. Diese GANs werden unter anderem zur Erstellung fotorealistischer Bilder, zur Visualisierung verschiedener Gegenstände, zur Modellierung von Bewegungsmustern in Videos und zur Erstellung von 3D-Modellen von Objekten aus 2D-Bildern verwendet. GANs werden auch zur natürlichen Gestaltung der Nutzerinteraktion mit Chatbots eingesetzt. GANs bieten eine Möglichkeit, auf Basis nicht vorklassifizierter Daten, neue Daten zu generieren, wodurch es in den oben genannten Bereichen einfacher und schneller wird entsprechende Ergebnisse zu erzielen.

Antwort

- Regelmäßige und gezielte Weiterentwicklung der eingesetzten Werkzeuge zur Erkennung von Deepfakes, um präventive und reaktive Maßnahmen ergreifen zu können
- Zusammenarbeit mit Forschung und Entwicklung
- Rechtliche Abklärung
- Internationale Zusammenarbeit
- Implementierung eines „Frühwarnsystems“, das die starke Verbreitung von schädlichen Deepfakes automatisch erkennt und an Behörden meldet

4 Auswirkungen für Österreich

In Österreich ist die Bedrohung durch Deepfakes real. Das Parlament geht davon aus, dass Deepfakes mittlerweile tagtäglich passieren würden, da keine umfangreiche Software mehr zu ihrer Erstellung erforderlich sei.[10] Prof. Hany Farid (UC Berkeley) meinte in einem Vortrag, dass in den nächsten drei bis fünf Jahren Deepfakes erstellt werden könnten, die nicht mehr als solche erkennbar wären. Die Erkennungssoftware sei immer einen Schritt hinterher.[11] Österreich entwickelt sich immer rascher in Richtung einer digitalen Gesellschaft, wodurch die Bedrohung durch solche Inhalte steigt. Deepfakes bergen Risiken für die nationale und internationale Sicherheit.

Ethische, demokratiepolitische sowie grund- und menschenrechtliche Aspekte müssen gewahrt werden. Die entsprechende internationale Zusammenarbeit auf EU- und internationaler Ebene ist sehr wichtig.

Die staatlichen Stellen werden eng und partnerschaftlich mit dem privaten Sektor zusammenarbeiten, insb. wird eine Zusammenarbeit mit Medien, Forschungseinrichtungen und der Privatwirtschaft eine wichtige Rolle im Kampf gegen Deepfakes sein. Der Bereich der Sensibilisierung für das Phänomen Deepfake muss ausgebaut werden, insbesondere hinsichtlich Medienkompetenzen für die Bevölkerung.

Die nationale KI-Strategie verweist u.a. auch auf zukünftige Herausforderungen im Zusammenhang mit Deepfakes.

Auch auf europäischer Ebene gibt es gegenwärtig verschiedene politische Ansätze und rechtliche Rahmenwerke, die sich mit dem Phänomen Deepfakes auseinandersetzen:

- Verordnungsvorschlag zu Künstlicher Intelligenz
- Datenschutzgrundverordnung
- Bestimmungen zum Urheberrecht
- E-Commerce Richtlinie
- Verordnungsvorschlag über einen Binnenmarkt für digitale Dienste
- Richtlinie über audiovisuelle Mediendienste
- EU-Verhaltenskodex zur Bekämpfung von Desinformation
- Aktionsplan gegen Desinformation
- Europäischer Aktionsplan für Demokratie

Das BMI wird besonders im Bereich der Bekämpfung von Cyberkriminalität/Sicherheit von Wahlen und der Sicherheitspolitik im Zusammenhang mit Deepfakes gefordert sein.

[10] APA Parlament, „Nationalrat fordert Strategie gegen Deepfakes“, 14.Oktober 2020 https://www.ots.at/presseaussendung/OTS_20201014_OTS0265/nationalrat-fordert-strategie-gegen-Deepfakes

[11] Vgl. High Level Videokonferenz zu Hybriden Bedrohungen des deutschen EU-Ratsvorsitz vom 19.11.2020

5 Ziele

- Sicherstellung des freien und unabhängigen Informationsbezugs für die österreichische Bevölkerung sowie der Grund- und Persönlichkeitsrechte
- Konsequenter Schutz der Integrität unserer Demokratie und der demokratischen Willensbildung vor Einflussnahme von außen
- Berücksichtigung ethischer, demokratiepolitischer sowie grund- und menschenrechtlicher Aspekte (einschließlich Kinder- und Jugendschutz)
- Schutz der verfassungsmäßigen Einrichtungen und der kritischen Infrastruktur Österreichs
- Stärkung der Widerstandsfähigkeit gegen Deepfakes auf Verwaltungsebene
- Bewusstseinsbildung und Stärkung der Kompetenz zur Verifikation digitaler Inhalte in der Bevölkerung
- Strafverfolgungskompetenzen evaluieren und gegebenenfalls ausbauen
- Zusammenarbeit in PPP-Modellen bzw. im Rahmen von Kooperationen mit der Wirtschaft
- Verbesserung der internationalen Zusammenarbeit auf EU- und internationaler Ebene

6 Maßnahmen

6.1 Bereits gesetzte Maßnahmen

- Im Auftrag des BKA vom AIT durchgeführte Studie zum Thema Resilienz gegen Desinformation
- KIRAS Forschungsprojekt „Defalsif-AI“[12] im Bereich Medienforensik und Desinformation gemeinsam mit dem Austrian Institute of Technology, der APA, dem ORF, u.a.
- Aktive europäische Zusammenarbeit
- Verstärkte interministerielle Zusammenarbeit

6.2 Geplante Maßnahmen

Man kann hier vier Handlungsfelder unterscheiden, die notwendig sein werden, um das Thema Deepfake umfassend zu behandeln.

1. Strukturen und Prozesse
2. Governance
3. Forschung und Entwicklung
4. Internationale Zusammenarbeit

Handlungsfeld 1 Strukturen und Prozesse

- Zur Umsetzung der Entschließung des Nationalrates vom 14.10.2020 und des Ministerratsvortrages wird es notwendig sein, das Thema Deepfake im Rahmen einer Schwerpunktgruppe zum Thema Desinformation zu behandeln:
  - Analyse der Probleme und Identifikation von Lösungsansätzen
  - Strategisches Vorgehen insbesondere in den Bereichen: Schutz der Demokratie, Schutz des Individuums, nationale Sicherheit und technologische Entwicklungen
  - Bei Bedarf: Bildung von Unterarbeitsgruppen für eine vertiefte inhaltliche Herangehensweise

[12] Das Sicherheitsforschungsprojekt Defalsif-AI hat u.a. die Entwicklung/Beschaffung eines Softwaretools zur Detektion von Deepfakes zum Ziel. Das Projekt wurde Ende 2020 gestartet und läuft noch bis Ende 2022, es gibt demnach noch keine konkreten Ergebnisse. Im Fokus des Projektes steht die Entwicklung von Maßnahmen gegen strategisch motivierte Desinformation. Die inhaltlichen Forschungsschwerpunkte liegen auf audiovisueller Medienforensik sowie Textanalysen und deren multi-modaler Fusion unter Zuhilfenahme von Methoden der künstlichen Intelligenz.

- Vorkehrungen auf behördlicher Ebene
  - Bündelung bestehender Kapazitäten und Prozesse
  - Weitere Vertiefung der Zusammenarbeit und interministerielle Vernetzung
  - Erhöhung der Kapazitäten im Bereich Monitoring und Screening
  - Sensibilisierung der öffentlich Bediensteten für das Thema Deepfake
  - Abstimmung mit anderen Arbeitsgruppen (Hybride Bedrohung), um kontinuierliches Vorgehen sicherzustellen

Handlungsfeld 2 Governance

- Die Rolle, Zuständigkeiten und Kompetenzen von Staat und nicht-staatlichen Akteuren werden unter Berücksichtigung relevanter Grund- und Menschenrechte festgelegt und Rahmenbedingungen für die Zusammenarbeit aller Beteiligten geschaffen
- Verstärkte Kooperation mit und mehr Pflichten für Online-Plattformen auf nationaler und speziell auf europäischer Ebene
- Die Schwerpunktgruppe Desinformation wird Analysen zu unterschiedlichen Aspekten, wie der rechtlichen Grundlage, erstellen und in ihrem Tätigkeitsbericht festhalten
- Aktivitäten im Bereich Bewusstseinsschaffung in der österreichischen Bevölkerung sollen gesetzt werden (in Kooperation mit Desinformation) zur:
  - Verstärkung der Medienkompetenzvermittlung in Schulen und für Erwachsene
  - Sensibilisierung der Nutzer für einen kritischen Medien- und Informationskonsum - speziell online
  - Evaluierung von tauglichen Lösungsansätzen für eine bessere Verifizierbarkeit von Onlineinhalten durch die User selbst

Handlungsfeld 3 Forschung und Entwicklung

- Zur Bekämpfung der Bedrohungen durch Deepfake ist eine hohe technische Expertise erforderlich. Im Rahmen der nationalen und der EU-Sicherheitsforschungsprogramme soll das Thema ein zentraler Forschungsschwerpunkt sein
- Die relevanten Stakeholder in Verwaltung, Wirtschaft, Medien und Forschung werden in gemeinsamen Projekten die konzeptuellen Grundlagen und die technologischen Instrumente zur raschen Erkennung von Bedrohungen aus Deepfake in Österreich entwickeln
- Bereits bestehende Forschungsarbeiten, wie z. B. DefalsifAI, sollen weiter ausgebaut werden

Handlungsfeld 4 Internationale Zusammenarbeit

- Österreich wird weiter eine aktive Rolle in der internationalen Zusammenarbeit auf europäischer und internationaler Ebene spielen, insbesondere im Erfahrungsaustausch, bei der Erstellung internationaler Strategien, bei der Erarbeitung freiwilliger und rechtlich verbindlicher Regelungen, bei der Strafverfolgung sowie bei Kooperationsprojekten
- Die Aktivitäten auf europäischer Ebene (RAS, Democracy Action Plan) werden weiterhin unterstützt und Österreich wird hier eine aktive Rolle einnehmen. Aktiv sollen die gesetzten Ziele der Europäischen Kommission im Bereich der geplanten Maßnahmen des Europäischen Aktionsplans für Demokratie verfolgt werden

7 Literaturverzeichnis

Aicher in Rummel/Lukas, ABGB[4] § 16 ABGB

Alex Engler, “Fighting Deepfakes when detection fails”, Brookings, 14. November 2019, https://www.brookings.edu/research/fighting-Deepfakes-when-detection-fails/.

APA Parlament, „Nationalrat fordert Strategie gegen Deepfakes“, 14. Oktober 2020 https://www.ots.at/presseaussendung/OTS_20201014_OTS0265/nationalrat-fordert-strategie-gegen-Deepfakes.

Deeptrace, “The State of Deepfakes”, Landscape Threats and Impact”, September 2019.

ENISA, Threat Landscape 2020, 20. Oktober 2020, https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends.

EUROPOL, UNICRI, Trend Micro, „Malicious Uses and Abuses of Artificial Intelligence, 2020, https://www.trendmicro.com/de_de/about/newsroom/press-releases/2020/2020-11-19-kriminelle-nutzen-kunstliche-intelligenz-nicht-nur-fur-Deepfakes.html.

ITA-AIT, „Deepfakes – Perfekt gefälschte Bilder und Videos“, November 2018 https://www.parlament.gv.at/ZUSD/PDF/FTA_03.pdf

Leo Kelion, “Deepfake detection tool unveiled by Microsoft”, BBC, 1. September 2020, https://www.bbc.com/news/technology-53984114.

Madeline Brady, “Deepfakes: a new disinformation threat”, 31. Juli 2020, Democracy Reporting International, https://democracy-reporting.org/de/dri_publications/Deepfakes-a-new-disinformation-threat/.

Meissel in Klang[3] § 16 ABGB.

Norbert Lossau, „Deepfake: Gefahren, Herausforderungen und Lösungswege“, KAS, Februar 2020, https://www.kas.de/de/analysen-und-argumente/detail/-/content/deep-fake-gefahren-herausforderungen-und-loesungswege .

Raphael Satter, “Experts: Spy used AI-generated face to connect with targets”, AP News, 13. Juni 2019, https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d .

Security Insider, „Mehr Desinformation mit Deepfakes und mehr Corona-Betrug“, 12. Jänner 2021, https://www.security-insider.de/mehr-desinformation-mit-Deepfakes-und-mehr-corona-betrug-a-990717/

Will Douglas Heaven, “Facebook just released a database of 100,000 deepfakes to teach AI how to spot them”, 2. Juni 2020, MIT Technology Review, https://www.technologyreview.com/2020/06/12/1003475/facebooks-deepfake-detection-challenge-neural-network-ai/.

REPUBLIK ÖSTERREICH Parlamentsdirektion SIEGEL	**Unterzeichner**	Parlamentsdirektion
	Datum/Zeit-UTC	2022-09-02T15:02:40+02:00
	Prüfinformation	Informationen zur Prüfung des elektronischen Siegels und des Ausdrucks finden Sie unter: https://www.parlament.gv.at/siegel
Hinweis	Dieses Dokument wurde elektronisch besiegelt.	