

Deepfakes

Regulatorische und institutionelle Handlungsempfehlungen



DEFAME FAKES

Mit der kommerziellen Verfügbarkeit generativer KI-Modelle zur synthetischen Bild-, Audio- und Videoerstellung wie Sora (OpenAI) oder Veo 3 (Google) hat die Verbreitung von künstlich erzeugten Medien und Deepfakes rasant zugenommen, speziell auf Social-Media-Plattformen. Generative KI-Anwendungen erlauben es, Personen, Orte und Ereignisse medial hochrealistisch zu imitieren und kohärent aufeinander abzustimmen, sodass die künstlich erzeugten Medien als echte Aufzeichnungen realer Begebenheiten erscheinen, obwohl sie sich nie ereignet haben.

Aufgrund dieser Realitätseffekte besitzen Deepfakes ein erhebliches **Täuschungspotenzial** und stellen ein wirkmächtiges Instrument dar, um die Wahrnehmung von Menschen zu beeinflussen und sie zu bestimmten Handlungen oder Denkweisen zu verleiten. Dieses Potenzial eröffnet feindseligen Akteur:innen vielfältige Möglichkeiten, Deepfakes für missbräuchliche Zwecke einzusetzen: Deepfakes heben die Glaubwürdigkeit von Betrugsmethoden und Manipulationspraktiken, steigern die Erzählfähigkeit von Desinformationskampagnen, erweitern die Mittel digitaler Diffamierung und eröffnen neue Angriffsvektoren, um Zugriff auf private oder institutionelle Konten und Sicherheitssysteme zu erlangen.

Um diesen Risiken durch Deepfakes entgegenzuwirken, wurden **Maßnahmen** formuliert, die darauf abzielen, die Verbreitung irreführender und betrügerischer Deepfakes einzudämmen, eine rasche Identifikation und Kennzeichnung synthetischer Medien zu ermöglichen, sowie Expertise, Unterstützung und Sicherheitsstandards zu etablieren. Als erforderlich erachtet werden insbesondere (1) verbindliche Pflichten für Online-Plattformen, denen bei der Verbreitung und Amplifikation irreführender und betrügerischer Inhalte eine maßgebliche Verantwortung zukommt, (2) eine rechtsverbindliche Festlegung von Mindeststandards für technische Kennzeichnung im Rahmen der Pflicht von Anbietern zur Ausgabe maschinenlesbarer Formate, (3) der Aufbau und die Bündelung forensischer Expertise, psychosozialer Unterstützung sowie zielgruppen-spezifischer Bildungsformate, um Behörden, Betroffene und Öffentlichkeit effizient zu unterstützen, sowie (4) die Implementierung von Sicherheitsstandards, speziell in sensiblen behördlichen Kommunikations- und Entscheidungskontexten, um Risiken vorzubeugen oder im Anlassfall rasch erforderliche Maßnahmen zu setzen.

I. Regulierung von Online-Plattformen

Online-Plattformen, insbesondere Social Media, stellen zentrale Infrastrukturen für die Verbreitung manipulativer, täuschender und betrügerischer Deepfake-Inhalte dar. Ein erheblicher Teil der durch Deepfakes verursachten Vermögensschäden entsteht durch die systematische **Bewerbung betrügerischer Produkte und Dienstleistungen** mittels bezahlter Werbeschaltungen. In der Praxis reagieren sehr große Online-Plattformen (VLOPs) auf entsprechende Hinweise häufig lediglich mit der Entfernung einzelner Anzeigen, während inhaltsgleiche oder sinngleiche Werbeschaltungen derselben Akteur:innen weiterhin verbreitet werden. Diese Praxis begünstigt eine fortgesetzte Täuschung von Nutzer:innen und unterminiert die Wirksamkeit bestehender Melde- und Abhilfemechanismen.

Da betrügerische Werbeanzeigen einen nicht unerheblichen Beitrag zu den Werbeeinnahmen großer Plattformen leisten, bestehen strukturelle ökonomische Anreize, einschlägige Inhalte nur selektiv oder verzögert zu entfernen. Vor diesem Hintergrund bedarf es einer regulatorischen Nachschärfung, um die konsequente und systematische Entfernung betrügerischer Inhalte sowie die dauerhafte Sperrung verantwortlicher Werbekonten sicherzustellen.

Eine der zentralen Herausforderungen im Kontext von Deepfakes ist zudem die rasche und wirksame **Rechtsdurchsetzung für Betroffene bildbasierter sexualisierter Gewalt**. Effektive Lösch- und Sperrmechanismen auf Seiten der Plattformbetreiber sind hierbei von besonderer Priorität, um schwerwiegenden Persönlichkeitsrechtsverletzungen zeitnah zu adressieren. In der Praxis erweisen sich jedoch sowohl Entfernung entsprechender Inhalte als auch die Identifikation und rechtliche Zuordnung verantwortlicher Akteure häufig als zeit- und kostenintensiv.

a

Strengere Bestimmungen zur Verifikation von Werbetreibenden: Empfohlen werden strengere Bestimmungen zur Verifikation von Werbetreibenden („Know-your-advertiser“), abgestuft nach Risikokategorie der beworbenen Produkte oder Dienstleistungen (insb. Finanz-, Investment- und Krypto-Werbung): Werbekonten sollen nur nach überprüfbarer Identitäts- und Unternehmensprüfung freigeschaltet werden; hierzu zählen insbesondere die verpflichtende (i) Angabe und Prüfung einer Unternehmenskennnummer (UID, VAT), (ii) nachvollziehbare Kontakt- und Unternehmenssitzdaten sowie ein (iii) Abgleich der Zahlungs- bzw. Kontoinformationen. Bei Verstößen oder wiederholten Auffälligkeiten sind Werbekonten zu sperren und eine erneute Registrierung zu untersagen.

b

Pflicht zur zeitnahen Abhilfe bei offensichtlichen Rechtswidrigkeiten: Empfohlen wird weiters eine Pflicht zur zeitnahen Abhilfe bei offensichtlicher Rechtswidrigkeit bei gemeldeten Inhalten durch akkreditierte Trusted Flagger. Diese umfasst die (i) unverzügliche Entfernung/Deaktivierung betrügerischer Werbeanzeigen, die (ii) Sperrung der zugehörigen Werbekonten sowie die (iii) Unterbindung identischer oder sinngleicher Werbeschaltungen, wie auch die (iv) zeitnahe Entfernung von Deepfakes, die Persönlichkeitsrechte verletzen, sowie sonstiger illegaler Inhalte. Erfolgt keine Entfernung, ist dies auf Betreiberseite nachvollziehbar zu begründen und zu dokumentieren.

c

Pflicht zur De-amplification potenziell betrügerischer und risikohafter Inhalte: Um die algorithmische Reichweite potenziell betrügerischer oder risikobehafteter Inhalte systematisch zu begrenzen, wird eine Pflicht zur De-amplification empfohlen. Auf diese Weise können ökonomische Fehlanreize zur Duldung oder Verstärkung betrügerischer Inhalte reduziert werden, ohne auf eine Löschpflicht zurückzugreifen.

d

Pflicht zur Unterbindung sinngleicher/nahezu identischer Re-Uploads: Um Rechtsverletzungen, fortgesetzte Täuschungshandlungen und erneute Schädigungen der Betroffenen zu verhindern, wird eine Pflicht zur Unterbindung sinngleicher/nahezu identischer Re-Uploads nach erfolgter Meldung und Entfernung (Active Monitoring in Bezug auf bereits identifizierte Inhalte) empfohlen.

e

Sanktionen bei Unterlassung gebotener Maßnahmen: Um strukturelle Anreize zur Duldung oder faktischen Begünstigung betrügerischer Inhalte regulatorisch zu korrigieren, sollten Sanktionen bei Unterlassen gebotener Maßnahmen veranlasst werden, sofern Plattformbetreiber trotz Kenntnis oder klarer Risikosignale erforderliche Schritte (z. B. De-amplification) nicht ergreifen.

f

Detektion synthetischer Profilbilder auf Dating-Plattformen: Empfohlen werden außerdem Maßnahmen zur Früherkennung von synthetischen Profilbildern auf Dating-Plattformen, insbesondere beim Hochladen von Profilbildern, um Identitätstäuschungen und Romance-Scams unter Einsatz synthetischer oder manipulierter Medien frühzeitig zu verhindern.

g

Beschleunigte Entfernung nicht-einvernehmlicher intimer Medien (NCIM): Um den Schaden für Betroffene sexualisierter digitaler Gewalt möglichst frühzeitig und wirksam einzudämmen, wird ein beschleunigtes Verfahren zur Entfernung nicht-einvernehmlicher intimer Medien (NCIM) bei offensichtlicher Rechtswidrigkeit nach Meldung durch Betroffene oder akkreditierte Stellen innerhalb einer 24-Stunden-Frist sowie die verpflichtende Unterbindung identischer Re-Uploads (unter Verwendung von Hash-Technologien) empfohlen.

II. Transparenz- und Kennzeichnungspflichten

Der AI-Act¹ reguliert Deepfakes primär über Transparenzpflichten für Betreiber (Offenlegung KI-erzeugter/manipulierter Inhalte) sowie über technische Kennzeichnungspflichten für Anbieter (Ausgabe maschinenlesbarer Formate)². Für die **Erstellung und Verbreitung von Deepfakes durch Privatpersonen** sind diese Pflichten jedoch grundsätzlich nicht anwendbar, da Nutzer:innen, die KI-Systeme ausschließlich im Rahmen persönlicher (nicht beruflicher) Tätigkeiten verwenden, nicht als Betreiber im Sinne des AI-Acts gelten.

Die **Pflicht von Anbietern zur Ausgabe maschinenlesbarer Formate** – etwa in Form von Wasserzeichen oder Metadaten – kann die Erkennbarkeit KI-erzeugter Medien zwar unterstützen, bietet jedoch keine Gewähr, da (i) entsprechende Kennzeichnungen im Lebenszyklus von Inhalten leicht verloren gehen oder entfernt werden können (z. B. durch Screenshots, Zuschnitt oder Re-Encoding), (ii) solche Kennzeichnungen nicht durchgängig von allen KI-basierten Mediengeneratoren implementiert werden und (iii) alternative oder unregulierte Generatoren diese Anforderungen vollständig umgehen können. Zudem verhindert eine rein anbieterseitige Kennzeichnung nicht, dass Deepfakes ungekennzeichnet auf Social-Media-Plattformen verbreitet werden.

Damit bleibt die **Eindämmung irreführender Deepfakes auf Social-Media-Plattformen** im Wesentlichen den Instrumenten des DSA³ überlassen (Risikominderung, Notice-and-Action). Der DSA verpflichtet sehr große Online-Plattformen und Suchmaschinen (VLOPs/VLOSEs) dazu, Maßnahmen zur Risikominderung zu treffen und sieht auch eine Kennzeichnung manipulierter Inhalte vor (Art 35 Abs I lit k DSA). Hierzu zählen unter anderem Risiken durch

die Verbreitung rechtswidriger Inhalte sowie tatsächliche oder absehbare negative Auswirkungen auf die Ausübung der Grundrechte, auf demokratische Prozesse und auf die öffentliche Sicherheit.

Die Durchführung der entsprechenden Prüfung obliegt zunächst den Plattformbetreibern, die im Rahmen ihrer DSA-Pflichten eigenständig bewerten müssen, ob Deepfake-Inhalte rechtswidrig oder als offensichtlich rechtswidrig einzustufen sind. Mangelnde Kennzeichnungen von Deepfakes sind jedoch nicht maßgeblich dafür, ob Inhalte rechtswidrig sind. Plattformen müssen Inhalte nicht allein wegen fehlender Kennzeichnung entfernen. Es ist insofern fraglich, ob und inwiefern der AI-Act und der DSA hinreichend wirksame Instrumente darstellen, um die Verbreitung ungekennzeichneter oder unzureichend gekennzeichnete Deepfakes auf Social-Media-Plattformen effektiv zu verhindern.

a

Pflicht zur menschenlesbarer Kennzeichnung synthetischer Medien: Dringend empfohlen wird eine Pflicht zur menschenlesbaren Kennzeichnung synthetischer Medien z. B. durch Labels wie „Made with AI“. Dies umfasst auch eine verpflichtende Kennzeichnung auf Social-Media-Plattformen, die nicht allein durch eine Selbstausskunft der User:innen abdeckbar ist.

b

Festlegung von technischen Mindeststandards der Anbieterkennzeichnung: Empfohlen wird eine rechtsverbindliche Festlegung von Mindeststandards für technische Kennzeichnung im Rahmen der Pflicht von Anbietern zur Ausgabe maschinenlesbarer Formate wie z. B. Wasserzeichen oder Metadaten, sodass Online-Plattformen technische Signale verlässlich auslesen können und synthetische Medien unterschiedlicher Anbieter nicht je eigene Ausleseverfahren erfordern (Interoperabilität). Empfehlenswert wäre außerdem, eine möglichst hohe Robustheit der Kennzeichnung über den Lebenszyklus der synthetischen Medien hinweg, sodass diese auch nach Re-Encoding oder Resizing erhalten bleiben.

¹Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz).

²Deepfakes müssen nach dem AI-Act durch Anbieter und Betreiber von KI-Systemen gekennzeichnet werden. Anbieter von KI-Systemen sind verpflichtet, dass Inhalte in maschinenlesbarem Format (z. B. Wasserzeichen, Metadaten) ausgegeben werden (Art 50 Abs 2 Satz 1 KI-VO) und Kennzeichnungen als technische Information enthalten sind, um synthetische Medien automatisiert identifizieren zu können. Betreiber müssen künstlich erzeugte oder manipulierte Medien durch Kennzeichnungen klar und deutlich für Nutzer:innen offenlegen (Art 50 Abs 4 Satz 1 KI-VO iVm ErwG 134 Satz 1 KI-VO).

³Digital Services Act, Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG (Gesetz über digitale Dienste).

III. Kompetenzaufbau, Support & Bewusstseinsbildung

Der missbräuchliche Einsatz synthetisch erzeugter Medien begründet einen strukturellen Bedarf an (i) forensischer Expertise zur Erkennung von Deepfakes, an (ii) fachübergreifenden Unterstützungsangeboten für Betroffene sowie an (iii) präventiven Maßnahmen zur Stärkung der Medien- und Informationsbewertungskompetenz. Dieser Bedarf betrifft insbesondere staatliche Stellen (Strafverfolgung, Verwaltung und Justiz), Medienorganisationen sowie betroffene Privatpersonen und Organisationen. Zur Deckung dieses Bedarfs wird die Einrichtung und/oder der Ausbau zentraler Anlaufstellen, koordinierter Austauschplattformen sowie die Bündelung einschlägiger Expertise empfohlen.

a

Einrichtung einer forensischen Fachstelle als behördliche Clearingstelle: Um spezialisierte forensische Expertise bereitzustellen, insbesondere für Strafverfolgungs- und Sicherheitsbehörden (einschließlich Staatsschutz und Nachrichtendiensten), Gerichte sowie für außenpolitische Institutionen, wird die Einrichtung einer forensischen Fachstelle empfohlen, die standardisierte fachliche Bewertungen potenzieller Deepfakes vornimmt. Durch Bündelung einschlägiger Kompetenzen und eine organisatorische Anbindung an bestehende Strukturen (z. B. Cyber Crime Competence Center, C4), kann eine ressourcenschonende und behördenübergreifende Bearbeitung von synthetischem Medienmaterial sichergestellt werden. Empfehlenswert erscheint in diesem Kontext auch eine Anbindung an bestehende EU-weite Initiativen oder Projekte (z. B. das EU-Projekt DETECTOR), um Ressourcen europaweit zu bündeln und spezialisierte Werkzeuge zur Identifizierung manipulierter Medieninhalte bereitzustellen.

b

Einrichtung/Ausbau einer Melde-, Informations-, und Auskunftsstelle: Zur systematischen Erfassung missbräuchlicher Vorfälle im Zusammenhang mit Deepfakes sowie zur Unterstützung von Betroffenen wird die Einrichtung und/oder der Ausbau einer Melde-, Informations- und Auskunftsstelle empfohlen. Diese soll (i) niedrigschwellige Unterstützungsangebote (technisch, rechtlich, psychosozial) bündeln und bereitstellen, (ii) standardisierte Abläufe für Dokumentation und Beweissicherung vorgeben und (iii) ein laufendes Lagebild zu Deepfake-bezogenen Vorfällen erstellen. Für Fälle bildbasierter sexualisierter Gewalt ist neben psychosozialer Beratung insbesondere technisch-operativer Support prioritär auszugestalten, um die rasche Entfernung bzw. Sperrung rechtsverletzender Inhalte zu erleichtern. Für einen effizienten und koordinierten Umgang mit missbräuchlichem Einsatz von Deepfakes werden zudem periodische Vernetzungstreffen relevanter Stakeholder, regelmäßiger **Wissensaustausch** sowie bedarfsgerechte **Weiterbildungsangebote** für Fachpersonal empfohlen. Dies trägt zu einer besseren Abstimmung der beteiligten Akteure, verkürzten Reaktionszeiten sowie erhöhter Handlungssicherheit bei.

c

Einrichtung eines forensischen Ansprech- und Kooperationspartners für Medien: Zur Qualitätssicherung journalistischer Berichterstattung und Wahrung einer faktenbasierten Meinungsbildung wird die Einrichtung eines forensischen Ansprech- und Kooperationspartners für Medienhäuser zur Verifikation von fragwürdigem Medienmaterial empfohlen (z. B. APA), um Redaktionen personell und finanziell zu entlasten. Die gezielte Verbreitung von Deepfakes zur Desinformation birgt das Potenzial gegen-aufklärerische Dynamiken und gesellschaftspolitische Konfliktlinien erheblich zu verstärken. Medienredaktionen können selbst zur Weiterverbreitung von Falschinformation beitragen, da Detektionssysteme keine vollständige Sicherheit liefern und auch Open-Source-Daten (OSINT) von Manipulation betroffen sein können.

d

Stärkung von Medien- und Informationskompetenz: Empfohlen wird zudem die gesamtgesellschaftliche Stärkung von Medien- und Informationskompetenz durch bildungsbezogene Maßnahmen, staatliche Informationsangebote und zielgruppenspezifische Aufklärungskampagnen, die sowohl präventiv einen verantwortungsvollen Umgang mit dem Erstellen und Teilen von Bild-, Audio- und Videomaterial fördern (Reduktion der „Angriffsfläche“) als auch Betroffene über bestehende Rechte informieren. Unabdingbar ist hierbei die zielgruppenspezifische Ausgestaltung der Bildungs- und Informationsmaßnahmen, da z. B. Jugendliche grundsätzlich andere Bedürfnisse haben als Personen in der nachberuflichen Lebensphase.

IV. Sicherheitsstandards

Die massentaugliche Erstellung und Verbreitung synthetischer Medien stellt ein noch junges technologisches Risikopotenzial dar. Dementsprechend bestehen aufseiten von Behörden, Gerichten und Exekutivorganen bislang keine standardisierten und institutionalisierten Handlungs- und Sicherheitsabläufe, um den Gefahren durch den gezielten Einsatz von Deepfakes präventiv vorzubeugen oder im Anlassfall rasch und konsistent erforderliche Maßnahmen zu setzen. Dies erhöht das Risiko von Fehlentscheidungen, unbefugten Zugriffen sowie der Weitergabe von sensiblen Informationen. Vor diesem Hintergrund wird empfohlen, verbindliche Sicherheitsstandards zu implementieren, die das Risiko durch synthetische Medien und Deepfakes in sensiblen Verwaltungs-, Ermittlungs- und Entscheidungsprozessen institutionell absichern und entsprechende Handlungsempfehlungen vorgeben.

a

Verpflichtende Zwei-Faktor-Authentifizierung: Da Deepfakes zur Umgehung biometrischer Authentifizierungssysteme eingesetzt werden können, stellen sie eine ernsthafte Bedrohung für elektronische Sicherheitsverfahren dar. Um unbefugte Zugriffe zu verhindern, wird für behördliche und administrative Zugangssysteme verpflichtende Zwei-Faktor-Authentifizierung empfohlen.

b

Sicherheitsmechanismen bei sensibler Kommunikation: Um Spionage- und Sabotageakte durch Identitätstäuschungen zu vermeiden, wird die Etablierung von Sicherheitsmechanismen bei sensibler Kommunikation (z. B. Rückrufverfahren, Mehrkanal-Verifikation) im Kontext von behördlichen Anweisungen, Zahlungsfreigaben oder sicherheitsrelevanten Entscheidungen empfohlen.

c

Forensische Schnelltests von Medienmaterial durch Detektionsverfahren: Um synthetische Medien frühzeitig in Verwaltungs- und Gerichtsprozessen zu identifizieren und zu berücksichtigen, werden forensische Schnelltests von Medienmaterial unter Rückgriff auf State-of-the-Art Detektionstools wie dem AIT Media-Intelligence-Tool empfohlen.

d

Regelung zum Einsatz synthetischer Medien als Beweismittel: Es wird empfohlen einen, verbindlichen Handlungsrahmen für den Einsatz synthetischer Medien als Beweismittel in Strafverfahren (z. B. zur virtuelle Rekonstruktion von Tathergängen) zu erstellen.

e

Regelungen von Krisen- und Wahlphasen als Hochrisikokontexte: Um den Schaden durch die Verbreitung von Deepfakes gering zu halten, wird die Einstufung von Krisen- und Wahlphasen als Hochrisiko-Kontexte durch behördliche Eilverfahren empfohlen, die bei Vorliegen plausibler Hinweise auf Desinformation oder Täuschungsabsicht eine priorisierte Prüfung fraglichen Materials auslösen, einen definierten Kommunikationskanal zu Plattformbetreibern vorsehen, um frühzeitig De-amplification-Maßnahmen zu veranlassen und eine beschleunigte Entfernung bzw. Deaktivierung durchzusetzen. Ergänzend sind klare Abläufe für behördliche Warnungen und Richtigstellungen festzulegen, um Fehlentscheidungen, Panikdynamiken und langfristige Vertrauensschäden wirksam zu minimieren.

Die Forschung in dieser Veröffentlichung wurde im Rahmen des Sicherheitsforschungsprogramm KIRAS des Bundesministeriums für Finanzen gefördert.



DEFAME FAKES

Kontakt

Julia Krickl
 krickl@oiat.at

 **Bundesministerium**
 Finanzen

Impressum

Österreichisches Institut für Angewandte
 Telekommunikation (ÖIAT)
 Ungargasse 64-66/3/404
 1030 Vienna, Austria



www.oiat.at
 office@oiat.at