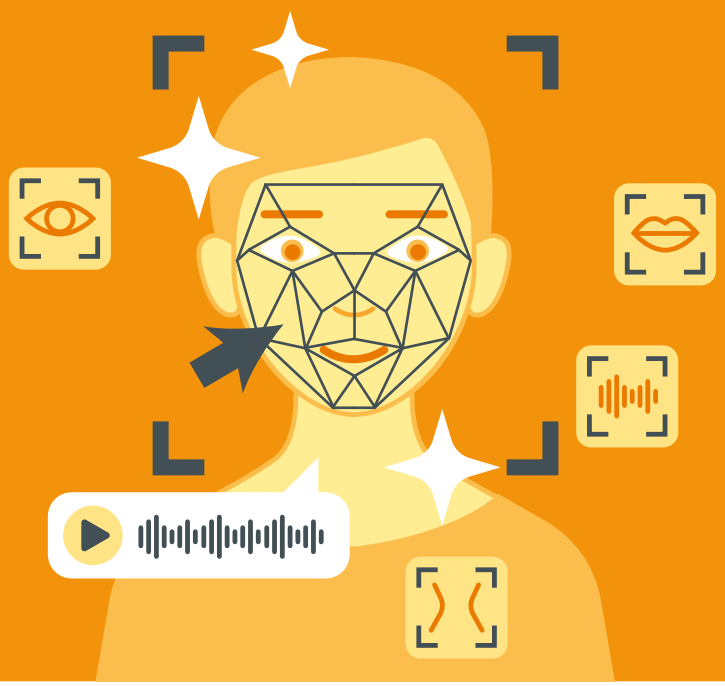
Saferinternet.at

Fachstelle digitaler Kinderschutz

FACT SHEET

Deepfakes und sexualisierte Gewalt

Was bedeutet das für den Kinderschutz im Internet?



Gefördert durch:

Bundeskanzleramt

Version: Oktober 2024

Fact Sheet

Deepfakes und sexualisierte Gewalt

Was bedeutet das für den Kinderschutz im Internet?

Künstliche Intelligenz ist in unserem Alltag angekommen. Das gilt auch für die zunehmend authentischer wirkenden Deepfakes, die auch die Lebenswelten von Kindern und Jugendlichen beeinflussen. Dieses Factsheet gibt einen Überblick über den Status quo von Deepfakes, zeigt aus einer Kinderschutzperspektive auf, welche Probleme Deepfakes im Hinblick auf (sexualisierte) Gewalt verursachen können, und bietet rechtliche Einordnungen sowie Tipps und Unterstützungsangebote.

1. Was sind Deepfakes?

Deepfakes setzt sich aus den Begriffen Deep Learning und Fakes zusammen. Deep Learning beschreibt dabei eine Methode der Künstlichen Intelligenz (KI), die komplexe Datenmuster analysiert und Fakes bedeutet nichts anderes als Täuschung.

Deepfakes sind also Medieninhalte (vor allem Audio-, Bild- und Videoaufnahmen), die mit Hilfe von KI erzeugt werden und realistisch wirken, obwohl sie es nicht sind. Deepfakes können beispielsweise Personen zeigen, die etwas sagen oder tun, was sie in Wirklichkeit nie gesagt oder getan haben.

Es gibt verschiedene Techniken, um Deepfakes zu erstellen. Grob kann zwischen manipulierten Inhalten (bestehende Fotos oder Bilder werden mit KI bearbeitet) und vollständig synthetischen Inhalten (vollständig von KI erstellte Bilder oder Videos) unterschieden werden:

Manipulierte Inhalte	
Face Swap	In einem Bild oder einem Video wird das Gesicht einer Person mit dem Gesicht einer anderen Person ausgetauscht. Mit entsprechender Hardware ist auch ein Echtzeit Face Swap möglich. **Tools:** DeepFaceLab, Faceswapper, Deepswap.ai etc. | **Beispiel:** ↗ https://www.youtube.com/watch?v=KE1LceuuxZg
Lipsync (Lippensynchronisation)	Eine neue Stimme wird über ein bestehendes Video gelegt und die Lippen werden synchronisiert. Diese Technik wird vor allem verwendet, um bekannten Personen Worte in den Mund zu legen, die sie nie gesagt haben. **Tools:** RunwayML Lipsync, Wav2Lip etc. | **Beispiel:** ↗ https://www.youtube.com/watch?v=IvY-Abd2FfM
Face Reenactment	In einem Video wird die Mimik einer Person auf das Gesicht der Person im Zielvideo übertragen. Mit leistungsfähiger Hardware ist auch ein Reenactment (dt. Nachahmung) in Echtzeit möglich. **Tool:** Face2Face etc. | **Beispiel:** ↗ https://www.youtube.com/watch?v=KUjn6SrNbSo

Face Morph	Bilder oder Videos von zwei verschiedenen Personen werden übereinandergelegt und so kombiniert, dass ein neues Gesicht entsteht. **Tools:** Face Morph, FaceShape etc. | **Beispiel:** ↗ https://www.youtube.com/watch?v=flxMYX8MGrg
Deep Nudes	Personen werden in Bildern oder Videos „entkleidet". Je nach Definition fallen darunter auch pornographische Face Swaps (Gesichter werden auf bestehenden pornographischen Darstellungen ausgetauscht). Schätzungen zufolge sind 98 % aller Deepfake-Videos im Internet pornografisch, 99 % davon betreffen Frauen.[1]
Vollsynthetische Inhalte	
Voice Cloning	Beim Voice Cloning wird eine Kopie einer menschlichen Stimme erstellt. Voice Cloning wird oft in Kombination mit Lippensynchronisation verwendet, so auch im dort aufgelisteten Beispiel. **Tools:** Speechify, ElevenLabs, Vocloner etc.
Generative KI-Bildgenerierung (Text-to-Image)	Auf Basis von realen Bildern wird eine KI trainiert, die mittels Texteingaben (Prompt) neue, noch nicht existierende Bilder erzeugt. **Tools:** Midjourney, Dall-E, RunwayML etc. | **Beispiele:** ↗ https://www.midjourney.com/showcase, http://thesecatsdonotexist.com/
Generative KI-Videogenerierung (Text-to-Video/ Image-to-Video)	Auf Basis von realen Videos wird eine KI trainiert, die neue, noch nicht existierende Videos erzeugt. **Tools:** Pika, RunwayML, Sora etc. | **Beispiele:** ↗ https://pika.art, https://www.youtube.com/watch?v=LKCtF0aE6uE

Aktuell handelt es sich bei den meisten Bild- oder Video-Fakes noch um sogenannte „cheap fakes" oder „shallow fakes". Dabei werden rudimentäre Bearbeitungstechniken (z. B. Photoshop-Bearbeitung und Videoschnitt) angewandt oder – noch einfacher – nicht gefälschtes Bild- oder Videomaterial aus dem eigentlichen Kontext gerissen, um neue Bedeutungszusammenhänge zu schaffen.

Deepfakes wirken zunehmend authentischer. Die Entwicklung neuer Algorithmen und leistungsfähigere Hardware beschleunigen und vereinfachen die Erstellung von Deepfakes, während die Qualität gleichzeitig besser wird. Damit wird auch das Erkennen von Deepfakes immer schwieriger.

2. Szenarien und strafrechtliche Einordnung

Im Folgenden gehen wir auf Szenarien ein, in denen es zu (sexualisierter) Gewalt gegen Kinder im strafrechtlichen Sinne kommt. Betroffene Personen dieser Szenarien können eine Strafanzeige bei der Polizei erstatten.

- **Cybergrooming**
- **Sextortion**
- **Cybermobbing**
- **Sexuelle Missbrauchsdarstellungen**
- **Beleidigende Inhalte**

Neben diesen strafrechtlich relevanten Aspekten kommt es in all diesen Szenarien in der Regel auch zu Verletzungen der Persönlichkeitsrechte und des Datenschutzrechts **(siehe Kapitel 3)**.

1 https://www.securityhero.io/state-of-deepfakes/

Cybergrooming

Beim Cybergrooming erschleichen sich (meist männliche) Erwachsene im Internet das Vertrauen von Kindern und Jugendlichen, um sie online sexuell zu belästigen oder um sich mit ihnen zu treffen und sie zu missbrauchen.

Rolle von Deepfakes: Um Vertrauen aufzubauen, geben sich die Täter als jemand anderes aus. Dazu benötigen sie Fake-Identitäten. Mit Bildgeneratoren können die Täter KI-Modelle so trainieren, dass zahlreiche Bilder oder Videos der gleichen Person erzeugt werden. Mit Hilfe von Echtzeitmanipulationen (z. B. Face Swap) sind sogar Videoanrufe möglich, bei denen sich die Täter als eine andere Person ausgeben können.

Rechtliche Einordnung:
Das Gesetz regelt in § 208a StGB („Anbahnung von Sexualkontakten zu Unmündigen") die Strafbarkeit von Cybergrooming. Danach ist bereits das Vorschlagen oder Vereinbaren eines persönlichen Treffens mit einer Person unter 14 Jahren über digitale Kommunikationswege strafbar. Gleiches gilt für die Kontaktaufnahme zu einer solchen Person mit der Absicht pornographische Darstellungen des Kindes zu erhalten. Ob dabei Deepfakes eingesetzt werden ist für den Straftatbestand irrelevant.

Sextortion

Sextortion bezeichnet einen Online-Betrug, bei dem Internetnutzer:innen von Unbekannten dazu aufgefordert werden, sich in einem Videochat auszuziehen oder sexuelle Handlungen an sich selbst vorzunehmen. Das Material wird heimlich aufgenommen, um die Betroffenen damit zu erpressen.

Rolle von Deepfakes: Bei Sextortion agieren die Täter mit falschen Identitäten. Deepfakes können daher auch bei Sextortion eingesetzt werden, um die für die Fake-Profile der Täter notwendigen Bilder und (Echtzeit-)Videos zu erzeugen. Außerdem verliert das heimliche Aufnehmen von Videos von den Opfern an Bedeutung. Denn Deepfake-Technologien ermöglichen es den Tätern, aus online verfügbarem Material selbst Deep Nudes der Betroffenen zu erstellen und sie damit zu erpressen – ganz ohne vorherige Kontaktaufnahme.

Rechtliche Einordnung:
Wird jemand unter Androhung der Veröffentlichung von intimen Aufnahmen oder Nacktaufnahmen zu etwas gezwungen (z. B. zum Übermitteln weiterer intimer Aufnahmen), stellt dies eine „Nötigung" gemäß § 105 StGB dar. Erpresst der Täter oder die Täterin die betroffene Person, um Geldleistungen zu erhalten, dann liegt eine „Erpressung" im Sinne des § 144 StGB vor. Soll die betroffene Person durch die Drohung gezielt in Angst und Unruhe versetzt werden, liegt eine weitere gerichtlich strafbare Handlung vor, nämlich eine „Gefährliche Drohung" gemäß § 107 StGB.

Cybermobbing

Cybermobbing ist das absichtliche Beleidigen, Bedrohen oder Bloßstellen einer Person über digitale Medien. Durch das Internet verbreiten sich solche Inhalte besonders schnell und sind nur schwer zu entfernen.

Rolle von Deepfakes: Deepfakes können verwendet werden, um Personen online zu mobben. Beispielsweise erstellen Jugendliche Deep Nudes von ihren Mitschülerinnen, um diese bloßzustellen. In einem Fall in Spanien wurden die Täter deshalb verurteilt.[2] Denkbar sind aber auch „harmlosere" Deepfakes, die z. B. Lügen und Gerüchte belegen sollen oder die Betroffenen in peinlichen Situationen zeigen. KI-gestützte Manipulationen wie Face Swap, Deep Nudes oder Lippensynchronisation sind für Jugendliche über zahlreiche Smartphone-Apps zugänglich.

Rechtliche Einordnung:
Cybermobbing kann nach § 107c StGB („Fortdauernde Belästigung im Wege einer Telekommunikation oder eines Computersystems") strafbar sein. Voraussetzung dafür ist, dass die Mobbinghandlung für eine größere Anzahl an Menschen (circa 10 Personen) über

2 https://futurezone.at/digital-life/ki-nacktbilder-deepfake-verurteilung-missbrauch-kinder-jugendliche-spanien/402923431

einen längeren Zeitraum wahrnehmbar gemacht wird. Die Mobbinghandlung muss darüber hinaus die betroffene Person in ihrer Lebensführung unzumutbar beeinträchtigen. Wenn die betroffene Person (bzw. eine andere Person an deren Stelle) sein Leben aufgrund des fortlaufenden Mobbings in wesentlichen Belangen ändert bzw. ändern würde (z. B. Schule oder Job wechseln, sich aus dem sozialen Leben zurückziehen), liegt eine solche unzumutbare Beeinträchtigung der Lebensführung vor.

Sexuelle Missbrauchsdarstellungen

Unter sexuellen Missbrauchsdarstellungen von Minderjährigen (Child Sexual Abuse Material, CSAM) werden wirklichkeitsnahe Abbildungen verstanden, die eine geschlechtliche Handlung an oder von Minderjährigen zeigen. Es fallen auch Darstellungen darunter, bei denen die Geschlechtsteile von Personen unter 18 Jahren im Fokus stehen und die der sexuellen Erregung des Betrachters dienen.

Rolle von Deepfakes: Die britische Internet Watch Foundation, die u.a. eine Meldestelle für sexuelle Missbrauchsdarstellungen von Minderjährigen betreibt, fand in einem Darknet-Forum innerhalb eines Monats mehr als 20.000 KI-generierte Bilder von Kindern, bei 3.000 davon handelte es sich um strafrechtlich relevante Missbrauchsdarstellungen.[3]

Dabei kommen vor allem Nudify- oder Face Swap-Methoden zum Einsatz (z. B. werden online verfügbare Bilder von Kindern in Nacktbilder umgewandelt oder Nacktbilder von Erwachsenen mit Kindergesichtern versehen). Auch Bilder von realen CSAM-Betroffenen oder prominenten Kindern werden verwendet, um KI-Modelle zu trainieren und Missbrauchsdarstellungen zu erstellen. Es gibt zahlreiche frei zugängliche Tools und Anwendungen, die auf diesen Methoden basieren. Altersbeschränkungen für die generierten Bilder und Videos fehlen häufig.

Rechtliche Einordnung:
Der Besitz, das Verbreiten, Herstellen, Zugänglichmachen sowie der wissentliche Zugriff auf Kindesmissbrauchsmaterial ist nach § 207a StGB („Bildliches sexualbezogenes Kindesmissbrauchsmaterial und bildliche sexualbezogene Darstellungen minderjähriger Personen") strafbar.

Darunter fallen nach § 207a Abs 4 Z 4 StGB auch Darstellungen, die vollkommen künstlich und/oder durch Manipulation von realen Bildern hergestellt wurden, wie es bei der Erstellung von CSAM mittels Deepfake-Technologien der Fall ist. Relevant ist dabei, dass diese den Eindruck vermitteln, dass es sich um eine Darstellung von einer tatsächlichen Missbrauchshandlung handelt.

Wird beispielsweise auf Basis einer Missbrauchsdarstellung einer realen betroffenen Person ein Deepfake erstellt bei dem deutlich erkennbar ist, dass es sich um eine künstliche Darstellung handelt, so ist dieses Bild rechtlich nicht als sexuelle Missbrauchsdarstellung zu werten. Wird hingegen von einer nicht existierenden Person ein fotorealistisches Fake von einer minderjährigen Person erstellt, kommt § 207a StGB zur Anwendung.

Wieso ermöglichen KI-Tools überhaupt die Erstellung von CSAM?
Damit dies möglich ist, müssen entsprechende Bilder oder Videos im Trainingsset der KI vorhanden sein. Dass dies tatsächlich der Fall ist, zeigte eine 2023 veröffentlichte Studie: Wissenschaftler:innen fanden Missbrauchsdarstellungen von Kindern in einem für Bildgeneratoren wichtigen Datensatz namens LAION-5B.[4]

Mit LAION-5B wurde unter anderem der beliebte Text-zu-Bild-Generator Stable Diffusion trainiert. In neueren Versionen von Stable Diffusion verhindern Sicherheitsmaßnahmen die

3 https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/
4 https://purl.stanford.edu/kh752sm9123

Generierung von CSAM. Als Open-Source-Datensatz waren die Daten jedoch frei zugänglich, wie oft der Datensatz für kleinere KI-Modelle verwendet wurde und wird, ist unbekannt.

Beleidigende Inhalte

Zusätzlich zu den bereits genannten strafrechtlichen Straftatbeständen kann bei beleidigenden Inhalten das strafrechtliche Privatanklagedelikt der „Üblen Nachrede" gemäß § 111 StGB oder der „Beleidigung" gemäß § 115 StGB begangen werden. Dies ist beispielsweise dann der Fall, wenn man mittels Deepfakes andere Personen bloßstellt, beleidigt oder verunglimpft.

Diese Privatanklagedelikte können nicht bei der Polizei angezeigt werden. Betroffene Personen können mit anwaltlicher Hilfe eine Privatanklage bei einem Strafgericht erheben und eventuell eine Entschädigung nach § 8 Mediengesetz fordern.

3. Persönlichkeitsrechte und Datenschutz

Die Erstellung von Deepfakes mit Aufnahmen des Gesichts, der Stimme oder des Körpers einer Person, verletzt deren Persönlichkeitsrechte und deren Datenschutzrecht. Sowohl die Personen, deren Originalmaterial verwendet wird, als auch die Personen, deren Gesicht, Körper oder Stimme für ein Deepfake verwendet wird, haben Unterlassungs- und Schadenersatzansprüche.

Persönlichkeitsrecht

Das „Recht am eigenen Bild" ist ein Persönlichkeitsrecht, das in § 78 UrhG geregelt wird. Demnach darf ein Foto oder Video einer Person nur unter bestimmten Voraussetzungen veröffentlicht oder verbreitet werden. Entscheidend ist, ob „berechtigte Interessen" der abgebildeten Person verletzt werden. Das ist zum Beispiel dann der Fall, wenn die abgebildete Person durch die Darstellung bloßgestellt wird (z. B. durch die Veröffentlichung eines Deep Nudes oder eines besonders peinlichen Deepfakes, **siehe Kapitel 2)**.

Ansprüche aus der Verletzung von Persönlichkeitsrechten können grundsätzlich nur auf dem zivilrechtlichen Weg, also durch eine Klage vor Gericht, durchgesetzt werden. Bei schwerwiegenden Persönlichkeitsverletzungen können Betroffene relativ rasch und kostengünstig im Wege des Mandatsverfahrens (§ 549 ZPO) vorgehen. Dazu muss lediglich ein Formular ausgefüllt und beim zuständigen Amtsgericht eingereicht werden.[5]

Datenschutzrecht

Beruhen Deepfakes auf Abbildungen einer real existierenden Person und ist diese Person erkennbar, liegt in der Regel auch eine rechtswidrige Datenverarbeitung und damit ein Verstoß gegen das Datenschutzrecht vor. Betroffene können eine Beschwerde bei der Datenschutzbehörde per Post oder E-Mail einbringen. Ein entsprechendes Formular ist auf der Website der österreichischen Datenschutzbehörde abrufbar.[6]

4. Tipps und Tricks zum Erkennen von Deepfakes

Waren Deepfakes bis vor kurzem in der Regel noch mit bloßem Auge zu erkennen, wird dies durch die rasant fortschreitende Professionalisierung von KI-Tools immer schwieriger. Kinder müssen deshalb lernen, skeptisch zu sein, wenn Aussagen oder Verhalten für eine bestimmte Person ungewöhnlich sind. Zusätzlich gibt es einige Tipps, mit denen viele der Deepfakes, die derzeit (noch) im Umlauf sind, erkannt werden können:

- **Unscharfe Übergänge:** Die Übergänge zwischen Gesicht und Haaren, zwischen Gesicht und Hals oder auch zum Hintergrund sind bei Deepfakes oft unscharf.

5 https://www.ombudsstelle.at/hass-im-netz/wie-funktioniert-das-mandatsverfahren-nach-549-zivilprozessordnung/
6 https://www.dsb.gv.at/download-links/dokumente.html

- **Unnatürliche Mimik und Bewegungen:** Bewegt sich die abgebildete Person ungewöhnlich, wirkt die Mimik des Gesichts unnatürlich oder der Blick leer, könnte es sich um ein Deepfake handeln.
- **Fehlendes Blinzeln:** Menschen blinzeln automatisch alle paar Sekunden, ohne sich dessen bewusst zu sein. Fehlt dieses Blinzeln in einem Video oder sieht das Blinzeln eigenartig aus, ist das ein Alarmsignal.
- **Unterschiedliche Qualitäten:** Vorsicht ist auch geboten, wenn das Gesicht eine andere Qualität als der Rest des Videos oder des Bildes hat.
- **Fehlerhafte Hände, Zähne oder Haare:** Bei Deepfakes sind Hände, Zähne oder Haare häufig fehlerhaft. Auf Unstimmigkeiten (bspw. sechs statt fünf Finger oder komisch herumfliegende Haare) an diesen Körperstellen sollte daher besonders geachtet werden.
- **Unpassende Stimme:** Bei Videos mit Ton sollte darauf geachtet werden, ob die Lippenbewegungen synchron sind, die Stimme nicht zur Person passt oder die Stimme sogar roboterhaft oder monoton klingt.

Hier bekommen Betroffene Unterstützung:

Rat auf Draht: Notruf für Kinder und Jugendliche – rund um die Uhr, anonym und kostenlos. Per Telefon (einfach 147 wählen) oder Chat:
↗ www.rataufdraht.at

Internet Ombudsstelle: unterstützt beim Entfernen von intimen Aufnahmen auf Social Media und berät bei rechtlichen Fragen mit Internetbezug:
↗ www.ombudsstelle.at

Stopline: Meldestelle gegen sexuelle Missbrauchsdarstellungen Minderjähriger & nationalsozialistische Wiederbetätigung im Internet: ↗ www.stopline.at

Polizei: Bei akuten Notfällen ist die nächstgelegene Polizeidienststelle unter der **Nummer 133** erreichbar. Für eine Strafanzeige kann die Polizeidienststelle auch direkt aufgesucht werden. Mitgenommen werden müssen ein amtlicher Lichtbildausweis und alle relevanten Unterlagen, die die Straftat belegen. Eine Strafanzeige kann bei jeder Polizeidienststelle oder Staatsanwaltschaft erstattet werden.

Gefördert durch:

Bundeskanzleramt