



Don't fake me

Hintergrundinformation

Künstliche Intelligenz ist aus unserem Alltag kaum noch wegzudenken. Doch was passiert, wenn KI nicht hilft, sondern Menschen schadet? Generative KI kann auf Grundlage erlernter Muster neue Texte, Bilder, Videos oder Audiodateien erzeugen. Dadurch können auch realistisch wirkende Darstellungen von Personen entstehen, die nie stattgefunden haben. Besonders problematisch ist die Nutzung dieser Technologie für sexualisierte Darstellungen ohne Zustimmung der Betroffenen.

Digitale sexuelle Gewalt

Digitale sexuelle Gewalt ist kein Randphänomen. Sie umfasst ein breites Spektrum an Handlungen: die nicht einvernehmliche Verbreitung intimer Bilder, die Erstellung und Verbreitung sexualisierter Deepfakes, oder sexualisierte Belästigung über soziale Netzwerke. Was all diese Formen verbindet: Sie greifen tief in die sexuelle Selbstbestimmung, die Würde und die psychische Gesundheit der Betroffenen ein. Trotzdem werden digitale Übergriffe häufig als „weniger real“ wahrgenommen und weniger ernst genommen.

Was sind Deepfakes?

Deepfakes sind mithilfe künstlicher Intelligenz manipulierte oder vollständig erzeugte Bilder, Videos oder Audiodateien, die reale Personen täuschend echt in Situationen darstellen können, in denen sie nie waren. Was früher aufwendig zu manipulieren war, kann heute mit wenigen technischen Schritten und teilweise auf Grundlage eines einzigen öffentlich zugänglichen Fotos erzeugt werden. Sexualisierte Deepfakes haben sich durch frei verfügbare KI-Bildgeneratoren und sogenannte „Nudification“-Angebote zu einer verbreiteten Form digitaler Gewalt entwickelt. Grundsätzlich kann damit jede Person zum Ziel werden – besonders häufig betroffen sind jedoch Frauen und Mädchen.

Fallbeispiel: Collien Fernandes

Der Fall Collien Fernandes macht sichtbar, welche Dimension digitale Gewalt durch sexualisierte Deepfakes annehmen kann. Fernandes hatte bereits vor Jahren öffentlich gemacht, dass KI-generierte sexualisierte Darstellungen von ihr im Internet kursierten. Der Fall zeigt, wie Deepfakes genutzt werden können, um Frauen gezielt zu sexualisieren, zu demütigen und einem potenziell unbegrenzten Publikum auszusetzen.

Fallbeispiel: Mädchen in Spanien

2023 wurden mehr als 30 Mädchen im Alter von 12 bis 14 Jahren aus der spanischen Stadt Almendralejo Opfer nicht einvernehmlich erstellter, KI-generierter Nacktbilder. Fotos der Mädchen aus sozialen Medien wurden mithilfe sogenannter „Nudify“-Apps manipuliert, die Kleidung auf Bildern künstlich entfernen

können. Die Bilder wurden anschließend über Messenger-Dienste verbreitet und in mindestens einem Fall zur Erpressung eingesetzt.

Fallbeispiel: Francesca Mani

Auch die damals 14-jährige Francesca Mani aus New Jersey wurde 2022 Opfer sexualisierter KI-Bilder. Mitschüler hatten pornografische Darstellungen von Schülerinnen erstellt; Mani gehörte zu mehr als 30 Mädchen, die möglicherweise betroffen waren. Der Fall zeigt, dass für die Erstellung überzeugender sexualisierter Deepfakes heute teilweise ein einziges Foto ausreicht. Damit sind nicht mehr nur Prominente gefährdet, sondern grundsätzlich alle Menschen, deren Bilder öffentlich zugänglich sind.

Verantwortung von Unternehmen und Staaten

Generative KI wirft deshalb nicht nur technische, sondern auch menschenrechtliche Fragen auf. Unternehmen tragen nach den UN-Leitprinzipien für Wirtschaft und Menschenrechte die Verantwortung, Risiken ihrer Technologien zu erkennen, zu verhindern und zu begrenzen. Dazu gehören menschenrechtliche Prüfungen bereits während der Entwicklung, Transparenz über die Erhebung und Verarbeitung von Trainingsdaten sowie wirksame technische Schutzmechanismen.

Auch Staaten stehen in der Verantwortung. Sie sollten unter anderem die rechtswidrige Sammlung personenbezogener Daten für KI-Training verhindern, Unternehmen zu menschenrechtlichen Folgenabschätzungen verpflichten, betroffene Gruppen in Regulierung und Entwicklung einbeziehen und leicht zugängliche Beschwerde- und Entschädigungsmechanismen schaffen.

Was macht die EU?

Im Juni 2024 trat die EU-Richtlinie zur Bekämpfung von Gewalt gegen Frauen und häuslicher Gewalt (Richtlinie 2024/1385) in Kraft. Die Richtlinie verpflichtet die Mitgliedstaaten, die nicht einvernehmliche Herstellung oder Verbreitung von intimen Bildern und manipuliertem Material, einschließlich sexualisierter Deepfakes, unter Strafe zu stellen. Die Umsetzungsfrist endet am 14. Juni 2027.

Besonders wichtig: Die Richtlinie verlangt auch, dass die zuständigen Behörden über die notwendigen Fachkenntnisse und Ermittlungsinstrumente verfügen. Polizei, Staatsanwaltschaften und Gerichte sollen für digitale Gewalt qualifiziert und entsprechend geschult werden. Dazu gehören Kenntnisse über die besonderen Formen digitaler Gewalt, ihre Auswirkungen auf Betroffene und die effektive Sicherung digitaler Beweise. Ziel ist auch, eine sekundäre Viktimisierung zu verhindern.

Ergänzt wird dieser Rahmen durch die EU-KI-Verordnung (Verordnung 2024/1689). Für Deepfakes gelten Transparenzpflichten: Künstlich erzeugte oder manipulierte Bild-, Audio- und Videoinhalte müssen grundsätzlich als solche erkennbar gemacht werden. Die Kennzeichnung allein verhindert jedoch weder die Erstellung noch die Verbreitung sexualisierter Deepfakes und kann den Eingriff in die Privatsphäre der Betroffenen nicht rückgängig machen.

Ab dem 2. Dezember 2026 sollen zudem bestimmte KI-Systeme verboten sein, die speziell zur Erzeugung nicht einvernehmlicher sexuell expliziter oder intimer Inhalte eingesetzt werden können. Dennoch bleiben Schutzlücken: Das Verbot richtet sich insbesondere gegen bestimmte KI-Systeme und bedeutet nicht automatisch, dass jede einzelne Handlung der Erstellung, Nutzung oder Verbreitung eines Deepfakes nach der KI-Verordnung verboten ist. Auch die private, nicht berufliche Nutzung ist teilweise vom Anwendungsbereich der KI-VO ausgenommen. Dadurch können bestimmte Pflichten, insbesondere Transparenzpflichten, bei der privaten Nutzung anders greifen als bei der professionellen Nutzung.

Was macht Deutschland?

In Deutschland ist die Erstellung sexualisierter Deepfakes bislang nicht als eigenständiger Straftatbestand umfassend geregelt. Der Referentenentwurf des Bundesministeriums der Justiz für ein Gesetz gegen digitale Gewalt (GdG) aus April 2026 soll Betroffene im Internet besser vor Hass, Verleumdung und digitaler Belästigung schützen und bestehende Schutzlücken bei bildbasierter digitaler Gewalt schließen. Dies betrifft auch nicht einvernehmlich erstellte und verbreitete sexualisierte Deepfakes.

Das Gesetzgebungsverfahren befindet sich derzeit in Vorbereitung: Nach der Vorstellung des Entwurfs durch Bundesjustizministerin Stefanie Hubig am 17. April 2026 folgte die Beteiligung von Fachverbänden und Organisationen. Der Gesetzentwurf wird derzeit zwischen den Ministerien abgestimmt. Ein Kabinettsbeschluss ist vorläufig für den 16. September 2026 vorgesehen. Nach einem Kabinettsbeschluss würde der Entwurf als Regierungsentwurf in Bundestag und Bundesrat eingebracht und dort beraten. Dabei kann der Gesetzestext noch verändert werden.

Recht allein reicht nicht

Deutschland muss sexualisierte Deepfakes wirksam bekämpfen – nicht nur auf dem Papier. Das Gesetzgebungsverfahren hat durch den Fall Colleen Fernandes an Fahrt aufgenommen, diesen Schwung darf es jetzt nicht verlieren! Wir müssen weiterhin den Druck auf die deutsche Regierung aufrecht erhalten.

Polizei, Staatsanwaltschaften und Gerichte müssen auch über das notwendige technische und fachliche Wissen verfügen, um digitale Gewalt und insbesondere sexualisierte Deepfakes zu erkennen, Beweise zu sichern und Betroffene wirksam zu unterstützen. Notwendig sind daher regelmäßige, qualitätsgesicherte und praxisnahe Schulungen. Diese sollten unter anderem vermitteln:

- wie Deepfakes und andere Formen digitaler Gewalt technisch funktionieren,
- wie digitale Beweise erkannt und gesichert werden,
- welche rechtlichen Möglichkeiten zum Schutz der Betroffenen bestehen,
- welche besonderen psychischen und sozialen Folgen digitale Gewalt haben kann

Die EU-Gewaltschutzrichtlinie verpflichtet Deutschland zudem, die zuständigen Behörden mit der notwendigen Expertise und den erforderlichen Ermittlungsinstrumenten auszustatten. Schulungen sind daher kein freiwilliges Zusatzangebot, sondern ein wesentlicher Bestandteil eines wirksamen staatlichen Schutzes vor digitaler Gewalt.

Quellen:

Amnesty Bericht in englischer Sprache: Unlawful by Design: Exposing the human rights costs of generative AI <https://www.amnesty.org/en/documents/pol40/0996/2026/en/>

<https://www.gwi-boell.de/de/2026/04/17/digitale-gewalt-ist-kein-einzelfall-der-fall-collien-fernandes-und-was-daraus-folgen>

<https://www.technologyreview.com/2023/12/01/1084164/deepfake-porn-scandal-pushing-us-lawmakers/>

<https://abcnews.com/US/mobile-apps-fueling-ai-generated-nudes-young-girls/story?id=103563734>

<https://www.bfdi.bund.de/DE/Fachthemen/Inhalte/Technik/KI-Verordnung.html>

<https://digital-strategy.ec.europa.eu/de/policies/regulatory-framework-ai>

<https://www.djb.de/presse/stellungnahmen/detail/st26-06>

<https://www.institut-fuer-menschenrechte.de/themen/geschlechtsspezifische-gewalt/eu-gewaltschutzrichtlinie>