klicksafe

Deep Fake. Deep Impact.

Wie Jugendliche Deepfakes erkennen und ihre Folgen kritisch hinterfragen lernen

Arbeitsmaterial für den Unterricht

Kofinanziert von der Europäischen Union

Titel:
Deep Fake. Deep Impact.
Wie Jugendliche Deepfakes erkennen
und ihre Folgen kritisch hinterfragen lernen

Arbeitsmaterial für den Unterricht

1. Auflage Mai 2025

Autorinnen:
Dominique Facciorusso (klicksafe), Sachteil und Projekte
Stefanie Rack (klicksafe) und jugendschutz.net, Projekte

Lektorat und inhaltliche Beratung:
Prof. Dr. Katharina A. Zweig (RPTU Kaiserslautern-Landau)
Prof. Dr. Justus Thies (TU Darmstadt)

Illustration:
Nele Konopka (nelekonopka.net)

Gestaltung und Layout:
Annette Lehmann (klicksafe)

Verantwortlich im Sinne des Presserechts:
Deborah Woldemichael (Projektleitung klicksafe)

Weitere Materialien von klicksafe gibt es unter:
www.klicksafe.de/materialien

Herausgeberin:
klicksafe
Medienanstalt Rheinland-Pfalz
Turmstraße 10
D-67059 Ludwigshafen
Tel.: +49 621 5202-271
info@klicksafe.de
www.klicksafe.de

klicksafe ist das deutsche Awareness Centre im Digital Europe Programme (DIGITAL) der Europäischen Union und wird von der Medienanstalt Rheinland-Pfalz verantwortet.

klicksafe ist Koordinatorin des Verbunds Safer Internet DE (www.saferinternet.de). Diesem gehören neben klicksafe die Internet-Hotlines internetbeschwerdestelle.de (durchgeführt von eco und FSM) und jugendschutz.net sowie die Helpline Nummer gegen Kummer an.

The project is co-funded by the European Union,
https://digital-strategy.ec.europa.eu/en/activities/digital-programme.

Die alleinige Verantwortung für diese Veröffentlichung liegt bei der Herausgeberin. Die Europäische Union haftet nicht für die Verwendung der darin enthaltenen Informationen.

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung Nicht kommerziell 4.0 International Lizenz, d. h. die nichtkommerzielle Nutzung und Verbreitung ist unter Angabe der Quelle klicksafe und der Webseite www.klicksafe.de erlaubt. Sollen über die genannte Lizenz hinausgehende Erlaubnisse gewährt werden, können Einzelabsprachen mit klicksafe getroffen werden. Wenden Sie sich dazu bitte an: info@klicksafe.de.

Weitere Informationen unter:
https://creativecommons.org/licenses/by-nc/4.0
Es wird darauf hingewiesen, dass alle Angaben trotz sorgfältiger Bearbeitung ohne Gewähr erfolgen und eine Haftung der Autor*innen ausgeschlossen ist.

Diese Broschüre wurde auf 100 % Recyclingpapier gedruckt.

Deep Fake. Deep Impact.

Wie Jugendliche Deepfakes erkennen
und ihre Folgen kritisch hinterfragen lernen

Arbeitsmaterial für den Unterricht

klicksafe

Inhaltsverzeichnis

Vorwort **5**

Deep… was?
Einführung zum Thema „Deepfakes“ **6**
Was sind Deepfakes? 6
Was ist Deep Learning? 6
Wie werden Deepfakes erstellt? 6
Was macht einen Deepfake „hochwertig“? 8

Welche Formen von Deepfakes gibt es? **9**
Textbasierte Deepfakes 9
Bildbasierte Deepfakes 9
Audiobasierte Deepfakes 13
Warum ist nicht jeder Fake „deep“? 14

Können Deepfakes sinnvoll sein? **16**
Unterhaltung 16
Kommerzieller Nutzen 18
Bildung 19
Empowerment 20
Gesundheit und Wohlbefinden 20

Können Deepfakes gefährlich sein? **21**
Cybermobbing 21
Hass und Hetze 21
Sexuelle Gewalt und Frauenhass 23
Missbrauch von Identitäten und Betrug 25
Desinformation und Demokratiegefährdung 25

Wie mit Deepfakes umgehen? **29**

Wie ist die Rechtslage? **30**

Weitere Informationen **31**

Übersicht über die Projekte **33**
Deepfakes auf der Spur –
Fakes verstehen, prüfen und erkennen 34
Die Macht der Bilder –
Risiken durch Deepfakes verstehen 38
Jetzt wird’s deep –
Deepfakes selbst erstellen 46

Quellenverzeichnis **51**

Hilfe und Beratung **54**

Vorwort

Was haben der ehemalige Papst Franziskus, die deutsche Politikerin Annalena Baerbock und der amerikanische Rapper Snoop Dogg gemeinsam?
Alle drei waren schon Gegenstand sogenannter Deepfakes, die in sozialen Netzwerken viral gingen. Dabei handelt es sich um manipulierte Medieninhalte, wie etwa Bilder oder Videos, die **mithilfe von künstlicher Intelligenz** (kurz: KI) verändert oder neu erstellt werden. Das heißt, man sieht oder hört etwas, dass es so nie gegeben hat. Wie etwa den vermeintlichen Papst in der hippen Daunenjacke, Frau Baerbock scheinbar in einem Pornofilm oder die mutmaßliche Stimme Snoop Doggs in einem Song, den er nie gesungen hat.
Die fortschreitenden Entwicklungen im Bereich der künstlichen Intelligenz haben Deepfakes zu einer Technologie emporgehoben, die uns fasziniert und neugierig macht. Gleichzeitig fordert sie uns als Gesellschaft aber auch zunehmend heraus, vor allem mit Blick auf unsere Urteilsfähigkeit und Medienkompetenz.
Deepfakes können beeindruckend und unterhaltsam sein, etwa wenn der 80-jährige Harrison Ford digital um die Hälfte verjüngt in einem neuen Film auf der Leinwand erscheint. Oder wenn der verstorbene Maler Salvador Dali den Museumsbesucher*innen sein Lebenswerk erzählt und am Ende fragt, ob sie vielleicht ein Selfie mit ihm machen möchten.
Deepfakes können auch witzig sein. Zum Beispiel wenn Olaf Scholz plötzlich zum tanzenden Tankstellenwart, Bodybuilder in der Muckibude oder Gangster-Rapper mutiert. Sie sind auch ein hilfreiches Tool, mit dem durch lippensynchrone Übersetzung in zahlreiche Sprachen Barrieren gesenkt und Teilhabe gefördert werden können.
Auf der anderen Seite wird die Technologie aber auch missbräuchlich genutzt. Mithilfe von KI lassen sich zum Beispiel Medieninhalte erzeugen, die Betroffene vermeintlich nackt oder bei sexuellen Handlungen zeigen. Insbesondere bei Frauen aus der Öffentlichkeit hört man immer wieder, dass KI-generierte Nacktbilder von ihnen online verbreitet werden und hohe Aufrufzahlen erreichen. Auch im Kontext politischer Desinformation wird immer wieder über Deepfakes gesprochen. Damit verbunden wird vor allem die Sorge, dass manipulierte Inhalte Bürger*innen in ihrem Wahlverhalten beeinflussen und den gesellschaftlichen Zusammenhalt schwächen bzw. sogar zerstören können.
Junge Menschen kommen mit Deepfakes in Social Media zunehmend in Berührung. Doch es ist nicht immer einfach, manipulierte Inhalte auf den ersten Blick zu erkennen. Oder zu verstehen, welche Motive und Absichten dahinterstecken (können). Die Sonderauswertung der PISA-Studie 2022 zeigt, dass sich die meisten Jugendlichen damit schwertun, die Qualität von Inhalten, die sie online finden, kritisch zu bewerten. Viele verlassen sich bei der Online-Recherche zum Beispiel auf nur eine Quelle oder prüfen Inhalte nicht auf ihre Korrektheit, bevor sie diese online mit anderen teilen.

Ziel des Materials ist daher, Lehrkräfte und Interessierten mit grundlegenden Sachinformationen zu versorgen, um sie und ihre Schüler*innen für das Thema „Deepfakes“ zu sensibilisieren und aufzuklären. Dabei geht es u. a. darum, was Deepfakes sind, wie sie erstellt werden und welche Formen es gibt. Es wird aufgezeigt, wie Deepfakes sinnvoll genutzt werden können, aber auch, welche Gefahren von ihnen ausgehen. Das Material bietet zudem zahlreiche Tipps und Informationen, wie man Deepfakes erkennen und mit ihnen umgehen kann. Zum anderen erhalten pädagogische Fachkräfte Projektideen für den Unterricht, um sich gemeinsam mit den Schüler*innen kritisch mit dem Thema auseinanderzusetzen und sie damit in ihrer Medien- und Informationskompetenz zu fördern.

Deep… was?
Einführung zum Thema „Deepfakes“

Was sind Deepfakes?

Bei Deepfakes handelt es sich um manipulierte Medieninhalte wie Texte, Audios, Bilder oder Videos, die mithilfe von künstlicher Intelligenz verändert oder neu erzeugt werden.

Der Begriff setzt sich aus den beiden Wörtern „Deep“ und „Fake“ zusammen. „Deep“, da hierfür **Deep Learning**, eine Form des maschinellen Lernens genutzt wird. Und „Fake“, da es sich bei den Inhalten um eine Fälschung bzw. einen computergenerierten Inhalt handelt. Je nach Qualität eines Deepfakes ist dieser manipulierte Inhalt nicht gleich als solcher zu erkennen oder kann **täuschend echt** wirken.
Der Begriff „Deepfakes“ wurde 2017 durch einen Reddit-Nutzer mit dem Benutzernamen „deepfakes“ geprägt. Dieser teilte in einem Unterforum des sozialen Netzwerks manipulierte Pornos, in denen die Gesichter der Darstellerinnen mit denen prominenter Frauen ausgetauscht wurden (sogenanntes Face Swapping). Andere Reddit-User nutzten das von ihm veröffentlichte Open-Source-Programm und erstellten eigene Deepfakes.
Im Verlauf der letzten Jahre hat sich die Leistungsfähigkeit von Deepfake-Programmen enorm weiterentwickelt. Zudem sind solche Tools über das Internet leicht zugänglich und können – je nach Anwendung – von Laien relativ einfach genutzt werden.

LESESTOFF

Das informative Comic-Essay „Schokoroboter & Deepfakes“ thematisiert KI aus der Perspektive von Jugendlichen.
→ **www.schokofakes.ai**

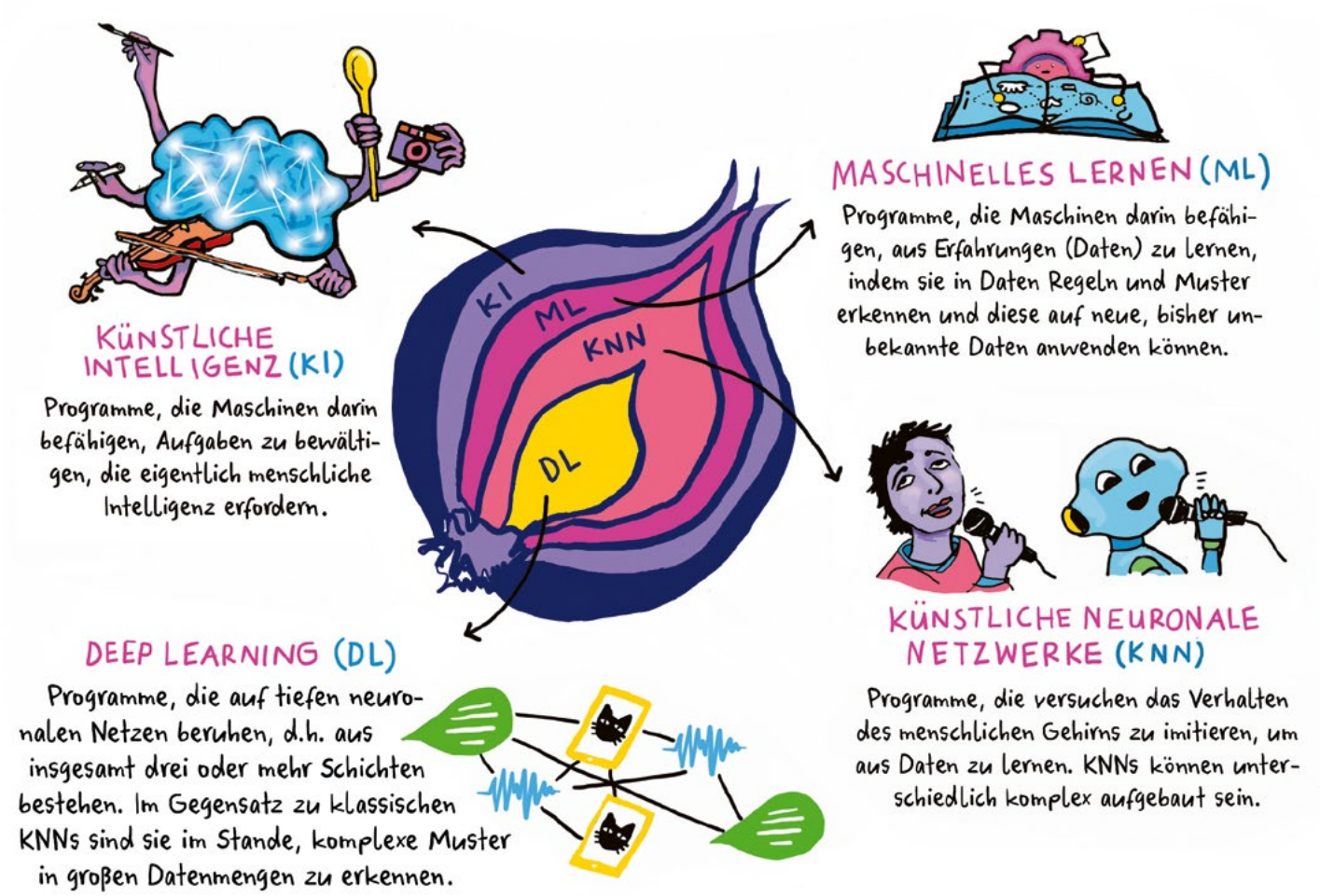
Im klicksafe Themenbereich „Künstliche Intelligenz“ wird unter anderem erklärt, wie maschinelles Lernen funktioniert.
→ **www.klicksafe.de/kuenstliche-intelligenz**

Was ist Deep Learning?

Maschinelles Lernen (kurz: ML) wird eingesetzt, um Computer mithilfe von Beispielen darauf zu trainieren, Muster und Zusammenhänge in Daten zu erkennen. Deep Learning (dt.: tiefes oder mehrschichtiges Lernen) ist ein wichtiger Teilbereich des maschinellen Lernens, mit dem sich viel komplexere Muster in Daten erkennen lassen, als dies mit klassischen ML-Modellen möglich ist. Das liegt daran, dass Deep Learning auf **tiefen künstlichen neuronalen Netzen** (kurz: KNN) basiert. Also Software, deren interne Struktur dem menschlichen Gehirn bzw. der Funktionsweise des Nervensystems nachempfunden ist. Aufgrund der neuronalen Netze lassen sich mit Deep Learning riesige Datenmengen verarbeiten, um komplexe Probleme zu lösen. Wie etwa bei der Sprach- und Bilderkennung.
Ein neuronales Netz ist immer aus verschiedenen nebeneinander gelagerten Schichten aufgebaut. Es gibt eine Ein- und Ausgabeschicht, um die Daten als Input in das System hinein- und später als Output wieder herauszugeben. Dazwischen liegt eine bzw. meist mehrere verborgene Zwischenschichten, auch „hidden layers“ genannt. Sie verleihen dem Netzwerk seine Tiefe. Jede Schicht enthält kleine Einheiten („Neuronen“ oder „Knoten“), die wie Gehirnzellen miteinander verbunden sind und Informationen weiterleiten.
Mit **„Learning“** ist gemeint, dass das Modell innerhalb dieser verborgenen Schicht(en) „lernt“, möglichst viele relevante Merkmale aus den Daten zu extrahieren und diese zu bewerten. Je mehr verborgene Schichten vorhanden sind, desto leistungsfähiger ist das Modell bzw. desto besser kann es komplexe Muster und Zusammenhänge in den Eingabedaten erkennen. Die Optimierung bzw. das Training des Modells erfolgt, indem die „Gewichtung“ der einzelnen Merkmale (Neuronen) so lange angepasst wird, bis das Ergebnis zufriedenstellend ist. Je mehr Daten vorhanden sind, desto besser ist das Ergebnis.

Wie werden Deepfakes erstellt?

Je nach Anwendungsbereich können unterschiedliche Technologien angewendet werden, um einen Deepfake zu erstellen. Die **drei am häufigsten genutzten Ansätze** sind Autoencoder, Generative Adversarial Network (kurz: GANs) und transformatorische Modelle. Alle Modellarten arbeiten mit künstlichen neuronalen Netzen.



Ihre **Ziele** und **Arbeitsweisen** sind jedoch **unterschiedlich** und sollen hier kurz skizziert werden:

1 Autoencoder: „Brühwürfel“ der wichtigsten Merkmale erzeugen

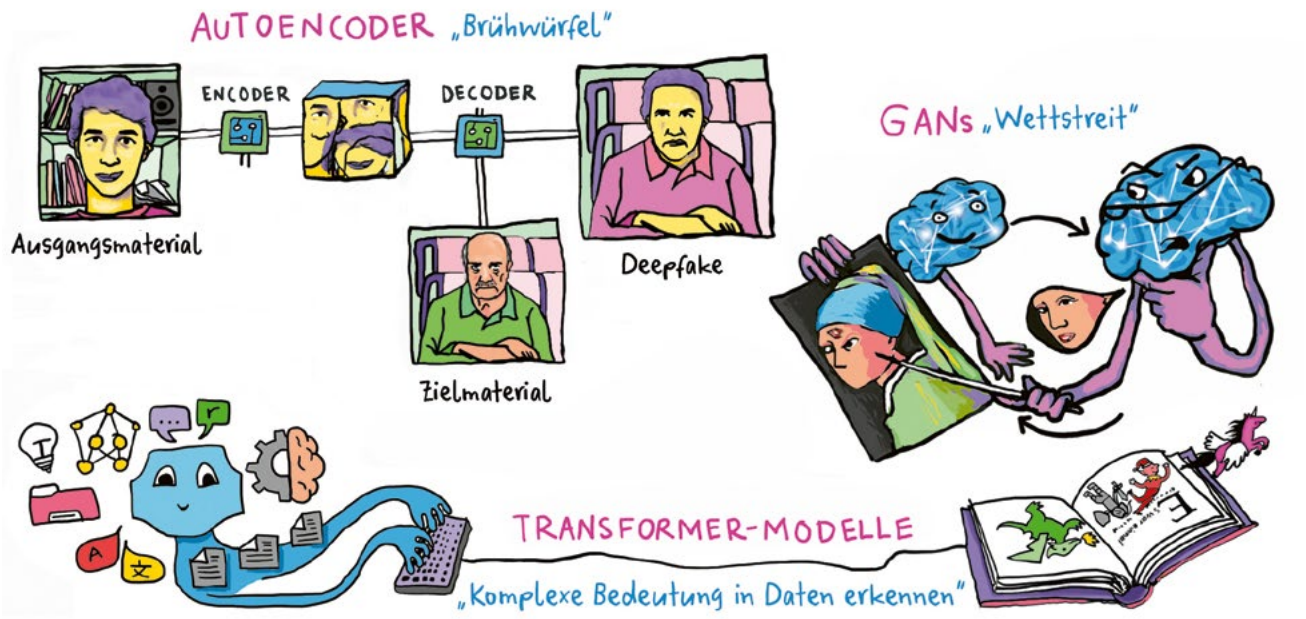
Autoencoder bestehen aus zwei Teilen, dem Encoder und Decoder. Will man einen bildbasierten Deepfake von einer **existierenden Person** generieren, benötigt man zunächst **Ausgangsmaterial** (z. B. Fotos der Person) für das Training des KI-Systems. Der **Encoder** hat die Aufgabe, diese Bilder zu analysieren und zu lernen, was die typischen Merkmale der abgebildeten Person sind (z. B. Form von Gesicht, Augen, Nase, Mund, etc.). Diese Erkenntnisse werden in einem kompakten Datenpaket abgespeichert. Es fasst die wichtigsten visuellen Eigenschaften der Person zusammen. Mithilfe dieses „Brühwürfels“ kann der **Decoder** die Person auf neue Weise rekonstruieren, also zum Beispiel ihre Mimik oder Bewegungen imitieren. Sollen nur bestimmte Teile der Person (z. B. Gesicht) in einen anderen Inhalt implementiert werden, wird auch ein **Zielmaterial** (z. B. bestehendes Video) benötigt.

2 GANs: Fälschungen durch „Wettstreit“ perfektionieren

Auch GANs bestehen aus zwei Teilen, dem Generator und Kritiker („Discrimator“). Die Modelle stehen in einer Art „Wettstreit“ zueinander, mit dem **Ziel, besonders realistische Fälschungen** zu generieren.

Mit GANs lassen sich **komplett neue Inhalte** generieren. Zum Beispiel Gesichter von **Personen, die es gar nicht gibt**. Um das umzusetzen, ist auch hier das Training mit echten Daten der erste Schritt. Denn die Modelle müssen zunächst lernen, wie echte Personen eigentlich aussehen. Danach beginnt der **Wettstreit**: Der **Generator kreiert** auf Basis dieser Erkenntnisse **neue Inhalte** (z. B. Gesichter) und der **Kritiker bewertet** diese Ergebnisse. Hierfür gleicht er sein antrainiertes Wissen über echte Gesichter mit den Fälschungen ab und gibt dem Generator **Feedback**, ob die Fakes echt wirken oder nicht. Der Generator versucht die nächste Fälschung unter Rücksichtnahme des Feedbacks zu verbessern. Die beiden Modelle lernen also voneinander. Dieser Prozess mündet darin, dass die **Qualität des Fakes zunehmend besser** wird und der Kritiker auf immer subtilere Fehler achten muss. Ziel ist es, eine so perfekte Fälschung zu erzeugen, dass der Kritiker sie nicht mehr erkennt.

Mit GANs lassen sich nicht nur neue Inhalte erzeugen, sondern auch **bestehende Daten optimieren**. Sie können etwa **in Kombination mit Autoencodern** zum Einsatz kommen, um die Qualität von erzeugten Deepfakes weiter zu verbessern.



3 Transformer-Modelle: Komplexe Bedeutung in Daten erkennen

Transformer-Modelle spielen eine **zentrale Rolle bei der Verarbeitung natürlicher Sprache**. Zum Beispiel um Texte zu übersetzen, zu klassifizieren oder zu generieren[1]. Ein bekanntes Beispiel ist der **Generative Pre-Trained Transformer** (kurz: GPT) von OpenAI[2]. „Generative" bedeutet, dass das Modell in der Lage ist, neue Inhalte zu erzeugen. Vorher muss das Modell mit sehr großen Datenmengen „vortrainiert" („Pre-Trained") werden, damit es die Strukturen und Regeln der Sprache erlernt. Der Vorteil der „Transformer"-Architektur besteht darin, **komplexe Beziehungen in Daten zu erkennen** (sogenannter Self-Attention Mechanismus). Transformer analysieren bei Texten etwa, in welcher Beziehung die Wörter zueinanderstehen und in welchen Kontext sie eingebettet sind.

Durch diese Fähigkeit eignen sich Transformer-Modelle besonders für die **Erzeugung zusammenhängender und menschenähnlicher Texte oder Sprachsequenzen**. Grammatik und Struktur sind meist korrekt und auch inhaltlich sind oder erscheinen sie sinnvoll. Solche Modelle können auch zur **Erstellung von text- und audiobasierten Deepfakes** genutzt werden. Zum Beispiel, um realistisch wirkende Aussagen zu erstellen, die nie gemacht wurden. Stellt man dem Modell personenbezogene Text- oder Audioproben zur Verfügung, kann es bei der Generierung sogar die **sprachlichen Merkmale einer bestimmten Person gezielt nachahmen**.

Was macht einen Deepfake „hochwertig"?

Jede Person kann theoretisch einen Deepfake erstellen. Online finden sich mittlerweile zahlreiche Deepfake-Tools, die sich ohne große Vorkenntnisse auf dem Smartphone nutzen lassen. Je nach App und Art des Inhalts kann die Qualität des Outputs jedoch stark schwanken.

Die **Qualität eines Fakes** hängt zum Beispiel von diesen Faktoren ab:

- **Komplexität des Modells:** Je mehr Parameter bzw. Merkmale ein Modell verarbeiten kann, desto detaillierter und präziser wird das Ergebnis. Sprich, wenn mehr Informationen berücksichtigt werden, kann das zu einer besseren Nachbildung der Realität führen.
- **Quantität der Trainingsdaten:** Je mehr Dateien (z. B. Bilder, Audios) von einer Person vorhanden sind, desto besser kann das Modell ihre biometrischen Parameter auslesen, um einen möglichst realistischen Fake von ihr zu erstellen.
- **Qualität der Trainingsdaten:** Je höher die Auflösung und Vielfalt der Eingabedaten, desto mehr Muster und Zusammenhänge können extrahiert werden. Wird ein Modell zum Beispiel mit Fotos trainiert, die eine Person in verschiedenen Perspektiven, Lichtverhältnissen und Kontexten zeigt, verbessert sich die Qualität des Ergebnisses. Personen aus der Öffentlichkeit eignen sich daher besonders für die Erstellung von Fakes, da es hier viel Ausgangsmaterial gibt.
- **Länge des Trainings und Rechenleistung des Computers:** Beides wirkt sich ebenfalls positiv auf die Qualität des Ergebnisses aus.
- **Nachbearbeitung:** Bei manipulierten Videos können viele verschiedene Perspektiven und (untypische) Bewegungen enthalten sein. Damit ein Fake natürlicher wirkt, ist es ggf. notwendig, einzelne Übergänge und Details anzupassen. Auch bei Audios kann dies der Fall sein.

Hochwertige Fakes werden noch nicht per Knopfdruck erstellt. Nahezu realistisch anmutende Fakes sind immer noch mit gewissen Skills, einer leistungsstarken Soft- und Hardware sowie Arbeit und Zeit verbunden. Nicht jede Person verfügt über solches Know-how bzw. die Ressourcen. Trotzdem wird es für Nutzende ohne technische Vorkenntnisse **immer leichter, „gut gemachte" Deepfakes zu erzeugen**.

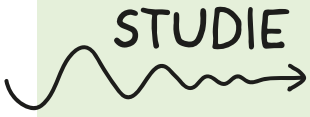
Welche Formen von Deepfakes gibt es?

Textbasierte Deepfakes

Auf Basis tiefer neuronaler Netze werden **große Sprachmodelle, sogenannte Large Language Modelle** (kurz: LLM), mit großen Textdatenbanken darauf trainiert, Texte in natürlicher Sprache zu verarbeiten und zu generieren. Leistungsstarke Chatbots sind zum Beispiel ChatGPT (OpenAI), DeepSeek R-1, Claude 3 (Anthropic) oder Gemini (Google). Durch sie lassen sich vielfältige Aufgaben umsetzen:[3]

- Sprachgeneratoren können **vorgegebene Texte bearbeiten**. Sie können sie zusammenfassen, in einfachen Worten erklären, in andere Sprachen übersetzen oder deren Inhalte analysieren und Fragen hierzu beantworten. Mit ihnen können auch der Inhalt, der Stil oder die Sprechweise des Textes verändert werden.
- Sprachgeneratoren können **neue Texte erstellen**. Das kann nützlich sein, um zum Beispiel eine Schreibblockade zu lösen, Vorschläge für den strukturellen Aufbau eines Referats zu erhalten oder ein Event zu planen, inklusive Zeitplan und Checklisten. Auch das ist in verschiedenen Sprachen möglich. Beim Erstellen der Antwort können Chatbots unterschiedliche Sprachstile imitieren, eine vorgegebene Rolle einnehmen oder den Inhalt an eine bestimmte Zielgruppe adressieren. Entscheidend ist der Befehl, auch „Eingabe“ oder „Prompt“ genannt, den man dem System hierfür gibt.
- Mit Sprachgeneratoren sind **Dialoge** möglich. Sie können eingegebene Inhalte interpretieren und den Kontext über einen längeren Gesprächsverlauf erfassen bzw. sich auf vergangene Inhalte beziehen. Sie sind auch im Stande, sentimentale Analysen durchzuführen. Also die Stimmung bzw. Emotionen des Nutzenden anhand der Eingaben zu erkennen.

Das Besondere an dieser neuen[4] Generation von Chatbots ist, dass sich die KI-generierten Texte lesen als wären sie menschlich. Der **Output wirkt eloquent, plausibel und damit vertrauenswürdig**. Sogar das **Nachahmen von Emotionen und Persönlichkeit** gelingt Chatbots zunehmend besser. Mittlerweile sogar als Audioausgabe einer KI-generierten Stimme (siehe GPT-4o[5]). All diese Fähigkeiten machen große Sprachmodelle zu leistungsstarken Werkzeugen, die sich sinnvoll, aber auch missbräuchlich nutzen lassen.



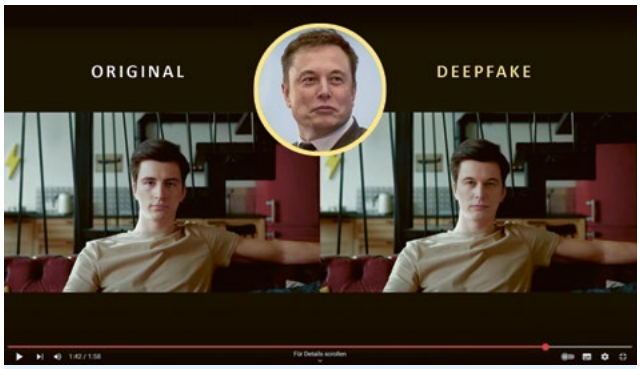
In einer **Studie (2024) der University of California** haben Wissenschaftler untersucht, ob Menschen erkennen können, wann sie mit einer Maschine kommunizieren. Grundlage hierfür war das Sprachmodell GPT-4 von OpenAI. In einer Online-Simulation führten die Teilnehmenden ein 5-minütiges Gespräch – entweder mit einem Menschen oder mit einem Chatbot. Anschließend sollten sie beurteilten, ob sie das Gesprächsgegenüber für menschlich hielten oder nicht. Das Ergebnis zeigt, dass **die Hälfte der Teilnehmenden den Chatbot fälschlicherweise für einen Menschen hielten**. Laut den Forschern deutet das Ergebnis darauf hin, dass Täuschungen mithilfe von KI-Systemen möglicherweise unentdeckt bleiben. Dass Chatbots Nutzenden ein Gefühl des „echten“ Austauschs geben können, zeigen auch die Ergebnisse der repräsentativen **Umfrage (2024) des Allensbach Instituts im Auftrag der Telekom**. Demnach haben **20 Prozent der Befragten**, die Chatbots regelmäßig nutzen, **schon mal vergessen, mit einer Maschine und nicht mit einem Menschen zu kommunizieren**.

Bildbasierte Deepfakes

Generative KI-Modelle können auch mit großen Mengen an Bilddaten trainiert werden, um visuelle Inhalte neu zu erzeugen bzw. zu verändern. Mit „neu“ ist gemeint, dass der Output etwas zeigt, dass es so nicht gibt. Trotzdem orientiert sich das Modell bei der Generierung eines visuellen Inhalts an anderen, gelernten Inhalten (Training des Modells). Ausgangsmaterial für das Training sind zum Beispiel Fotos, Grafiken oder Illustrationen. Bei der Erstellung bildbasierter Deepfakes werden **verschiedene KI-basierte Verfahren** genutzt. Mit ihnen lassen sich entweder Teilbereiche eines Bildes oder Videos verändern oder komplett neue Inhalte erstellen. Dies sind gängige Verfahren zur Erstellung von bildbasierten Deepfakes (in Anlehnung an Karaboger et al., 2024):

1 Gesichter austauschen (Face Swapping)

Gesichter lassen sich durch KI auf unterschiedliche Weise fälschen. Ein gängiges Verfahren, das seit der frühen Phase von Deepfakes zum Einsatz kommt, ist das sogenannte Face Swapping. Manchmal auch als „Face-Replacement“ bezeichnet. Bei dieser Bild- und Videobearbeitung wird das **Gesicht einer Person durch das einer anderen Person ersetzt**. Dabei



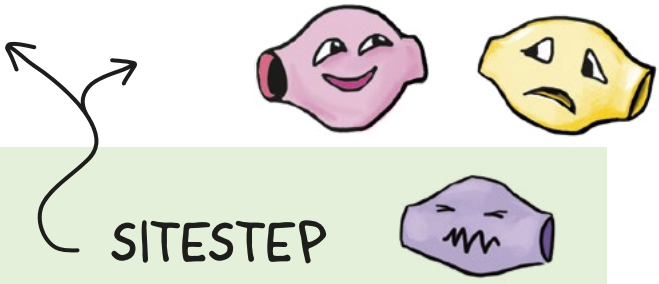
Face Swap
In dem Video wird gezeigt, wie sich mithilfe von KI Gesichter austauschen lassen. *Quelle: Screenshot, YouTube (2024). Demo: www.youtube.com/watch?v=wWS0dr9V4Ss*

Face Reenactment
Das Video zeigt, wie die Mimik und Lippenbewegung einer Zielperson („Target Actor“) in Echtzeit durch einen Manipulator („Source Actor“) gesteuert werden kann. *Quelle: Screenshot, https://justusthies.github.io/posts/face2face (2016). Demo: www.youtube.com/watch?v=ohmajJTcpNk*

erkennt die KI die Gesichtszüge, Mimik, Proportionen und Beleuchtung der beiden Gesichter und stimmt diese aufeinander ab. Ziel ist es, dass sich das eingefügte Gesicht möglichst nahtlos in ein anderes Bild oder Video integriert, ohne dass die Anzeichen der digitalen Manipulation ersichtlich sind.

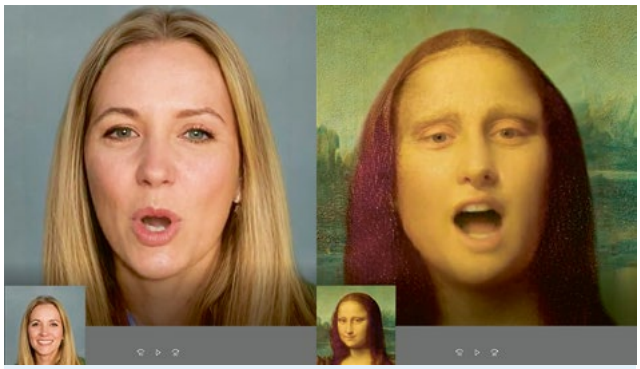
2 Gesichtsausdruck manipulieren (Face-Reenactment)

Ein weiteres Verfahren der Gesichtsfälschung ist das sogenannte Face-Reenactment, das hauptsächlich bei Videomaterial eingesetzt wird. Hier werden in einem bestehenden Video die **Mimik und Gesichtsbewegungen einer Zielperson** (Target Actor), wie die der Augen, Lippen oder des Kopfes, **an die einer anderen Person angepasst** (Source Actor). Die Identität der Zielperson bleibt jedoch erhalten. Der „Source Actor“ bewegt zum Beispiel seinen Mund (sagt etwas) oder bringt mit seiner Mimik etwas zum Ausdruck, das auf den „Target Actor“ rein visuell übertragen wird (siehe Abbildung). Es geht hier also erstmal nur um die Bewegung, die übertragen wird. Kombiniert man diese Manipulation noch mit einer Tonspur, kann man bei dem Deepfake auch hören, was der „Source Actor“ mit den Mundbewegungen „spricht“.
Der Zielperson können damit **Worte in den Mund gelegt werden, die sie nie gesagt hat und ihre Mimik ist wandelbar**. Mit manchen Face-Reenactment-Modellen ist sogar eine Übertragung der Mimik in Echtzeit möglich (sogenanntes Real-time-Reenactment). Dies lässt sich aber nicht nur auf natürliche Personen, sondern auch auf virtuelle Charaktere oder Avatare anwenden.
Ein weiteres Beispiel für Face-Reenactment ist das sogenannte **Lip Syncing** (dt.: Lippensynchronisation). Hier wird nur ein **bestimmter Bereich** des Gesichts mithilfe von KI verändert, nämlich die Bewegungen der Mund- und Kieferregion. Mit solchen Verfahren lassen sich Videos in andere Sprachen übersetzen (Dubbing), wobei sich die Lippen der sprechenden Person **synchron zur Tonspur** bewegen. Es ist mithilfe KI-erzeugter Stimm-Klone sogar möglich, dass die KI-generierte Übersetzung wie die Person im Original-Video klingt (mehr Infos unter „Audiobasierte Deepfakes“). Beispiele für solche Tools wären Synthesia und HeyGen.

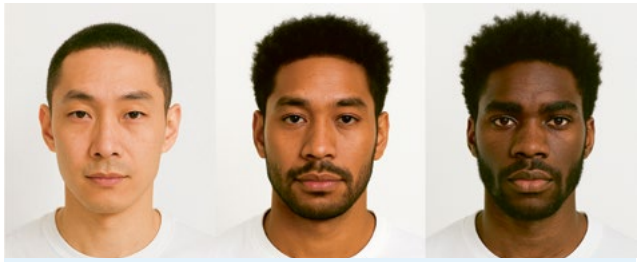


Ein Gesicht sagt mehr als 1000 Worte

Hast du gewusst, dass die Mimik einer der grundlegenden Bausteine der zwischenmenschlichen Kommunikation ist? Genau wie die Körpersprache und Gestik ist die Mimik ein wichtiger Bestandteil der „nonverbalen“ Kommunikation. Heißt: Kommunizieren, ohne mit Worten zu sprechen. Denn über unseren Gesichtsausdruck vermitteln wir dem Gegenüber wichtige Signale und unausgesprochene Botschaften. Bereits subtile Bewegungen zeigen, wie wir uns fühlen und ob das, was wir sagen, im Widerspruch dazu steht. Darüber erkennbar ist auch, welche Absichten wir im Sinn haben. Die Fähigkeit, nonverbale Signale zu erkennen und richtig zu interpretieren ist essenziell, um andere zu „entschlüsseln“. Das ist zum Beispiel wichtig, damit wir unser Gegenüber richtig verstehen und eine Beziehung zu der Person aufbauen können.



„Talking Face"-Generatoren
Wie realistisch sprechende KI-Avatare wirken können, zeigt Microsoft mit seinem Video-Generator „VASA-1". *Quelle: Screenshot, YouTube (2025). Demo: www.microsoft.com/en-us/research/project/vasa-1*
Auch der chinesische Konzern Alibaba haucht mit seinem KI-Tool „Emotive Portrait Alive" (kurz: EMO) statischen Bildern Leben ein.
Demo: www.youtube.com/watch?v=wtcSZdHZne4



Face Morphing
Links und rechts sind zwei verschiedene Personen zu sehen. Das Bild in der Mitte zeigt die Verschmelzung der beiden Identitäten. Alle Bilder wurden mit Hilfe von KI generiert. *Quelle: ChatGPT 4o, OpenAI (8.4.2025)*

Face Synthesis
Links: Das Bild wurde mithilfe von KI generiert.
Quelle: ChatGPT 4o, OpenAI (8.4.2025)
Rechts: Das KI-generierte Bild stammt von der Webseite „This Person does not exist".
Quelle: thispersondoesnotexist.com (2025)

Zu den aktuell bekanntesten Bildgeneratoren zählen: DALL-E (OpenAI), GPT-4o (OpenAI), Midjourney, Stable Diffusion und Adobe Firefly (Stand: Mai 2025).

TOOLS

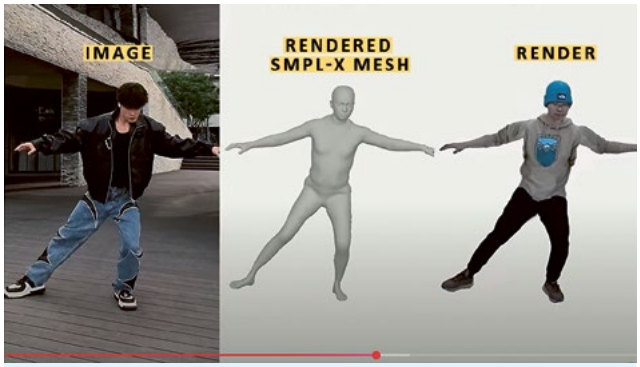
Mit Face-Reenactment-Tools lässt sich sogar die **Mimik von Personen in starren Bildern verändern**, indem man zum Beispiel Fotos animiert. Solche kurzen Videoclips mit lebensechten Bewegungen lassen sich zum Beispiel über My Heritage erstellen. Nach Angaben des Unternehmens geht es bei der Anwendung darum, verstorbene Menschen in Bewegtbilder umzuwandeln. Wie zum Beispiel historische Persönlichkeiten oder Angehörige. Kombiniert man solche animierten Portraitfotos dann noch mit Sprachaufnahmen, erhält man sprechende Gesichter. Diese **„Talking Faces"** wirken in ihrem Blick, ihrer Mimik und Kopfbewegung zunehmend natürlicher. Auch die Bewegungen von Mund und Kiefer werden immer synchroner zum Ton.

3 Gesichter miteinander verschmelzen (Face Morphing)

Das sogenannte Gesichts-Morphing ist ein weiteres Verfahren der KI-gestützten Manipulation von Gesichtern. Ziel ist es, die **Gesichtszüge mehrerer Personen miteinander zu verschmelzen**. Hierfür lernt die KI anhand von vorgegebenen Bildern (mindestens zwei), welche Merkmale die dort abgebildeten Personen haben. Auf dieser Basis wird ein neues Bild generiert, das jedoch Eigenschaften von beiden Individuen enthält. Sprich, ein „gemorphtes" Bild hat am Ende Ähnlichkeiten zu zwei oder mehreren Identitäten. Je nachdem wie viele Personen als Grundlage vorgegeben wurden.

4 Gesichter neu generieren (Face Synthesis)

Im Gegensatz zu den anderen Manipulationsarten geht es bei diesem Verfahren um die Synthese eines komplett neuen Gesichts. Ziel ist es, **realistisch aussehende Personen zu erzeugen, die es gar nicht gibt** (vgl. Kapitel „Wie werden Deepfakes erstellt?"). Die Qualität und Detailschärfe solcher KI-basierten Identitäten hat sich in den letzten Jahren extrem verbessert. Mittlerweile lassen sich solche künstlich erzeugten Identitäten auch als Bewegtbild generieren.
Ruft man die Webseite **„thispersondoesnotexist.org"** auf, erscheinen Bilder von Personen, die es gar nicht gibt. Aktualisiert man die Seite, erzeugt das dahinter liegende neuronale Netzwerk (GANs) jedes Mal einen neuen Deepfake. Das Besondere: Die Bilder sehen täuschend echt aus. Doch achtet man auf Details wie Ohren, Pupille (sie ist meistens nicht rund), Übergänge oder den Hintergrund, lassen sich die Fakes nach aktuellem Stand oft entlarven.



Body Puppetry
Mithilfe von KI lassen sich mittlerweile digitale Körper wie Marionetten durch die Bewegung anderer Körper („Source“) steuern. Das Video zeigt beispielsweise einen ausdrucksstarken 3D-Avatar, der auf Basis eines nur kurzen Videos erstellt wurde. Dabei wurden nicht nur die Körperbewegungen eines „Source Actors“ auf die Zielperson übertragen, sondern auch die Mimik. *Quelle: Screenshot, YouTube (2025).*
Demo: www.youtube.com/watch?v=GzXlAK-sBKY

5 Körpermarionette (Body Puppetry)

Mithilfe von KI lassen sich nicht nur die Bewegungen von Lippen, Mimik oder Kopf beeinflussen. Mit dem „Body Puppetry“-Verfahren werden **Bewegungen einzelner Körperteile bzw. des gesamten Körpers manipuliert**. So als würde man eine Marionette steuern. Man erzeugt einen Avatar, der entweder eine echte Person auf Grundlage eines Bildes oder Videos darstellt oder eine künstliche, rein virtuelle Figur.

6 Aus Text Videos generieren (Text-to-Video)

Mit Text-to-Video-Generatoren[6] lassen sich **Textbeschreibungen in Video-Inhalte** umwandeln. Der Generator analysiert dafür den Text („Prompt“) nach Informationen, die ihm vorgeben, was er genau darstellen soll. Ähnlich wie in einem Drehbuch. Text-to-Video-Generatoren gibt es inzwischen einige im Internet. Viele bieten zusätzliche Funktionen an, mit denen sich der **Stil, die Stimmung oder das Format** im Video anpassen lassen. Bei manchen Tools können auch akustische Inhalte, wie gesprochene Sprache, Musik oder Hintergrundgeräusche implementiert werden. Je nach Video-Generator unterscheiden sie sich auch mit Blick auf Kosten, Benutzerfreundlichkeit und Qualität.

Die **Qualität der KI-generierten Videos** hat sich in den letzten Jahren **enorm weiterentwickelt**. Besonders, wenn es um **realistisch anmutende Inhalte** geht, wie Menschen, Tiere, Landschaften oder Gesetze von Physik und Logik. Zu Beginn des Jahres 2024 sorgte **„Sora“**, das KI-Video-Tool von OpenAI, für viel Aufsehen. In einem offiziellen Launch veröffentlichte das Unternehmen kurze Video-Clips, die in punkto Detailschärfe und Fotorealismus ein völlig neues Niveau im Bereich der Videogenerierung zeigten. Der Zugang wurde anfangs auf einen ausgewählten Personenkreis begrenzt, um das Modell zu testen, Feedback zu erhalten und Sicherheitsrisiken zu klären. Seit Ende 2024 steht es auch zahlenden ChatGPT-Kunden in vielen Ländern zur Verfügung.
Seit der Verkündung von Sora haben auch andere Anbieter unter Hochdruck daran gearbeitet, eigene Video-Generatoren auf den Markt zu bringen. Zum Beispiel Google (Veo) und Meta (Movie Gen). Einige **Konkurrenz-Modelle** kommen zum Teil an die Qualität von Sora heran oder übertreffen sie sogar. Aktuelle Beispiele wären Runway mit seinem neuesten Modell „Gen-3 Alpha“ oder auch Konkurrenz-Modelle aus China wie „Kling AI“ (Kuaishou) und „Hunyuan“ (Tencent). Hunyuan generiert sogar Videos ohne inhaltliche Einschränkungen. Weitere bekannte Video-Generatoren sind Minimax, Luma, Pika Labs, Pictory, Synthesia oder Canva.

Text-zu-Video-Generatoren sind trotz beeindruckender Qualität **noch nicht vollständig ausgereift**. In der Umsetzung haben die Modelle zum Beispiel oft noch ein Problem mit **„Kohärenz“**. Damit ist die Fähigkeit gemeint, dass die Bilder über längere Zeiträume hinweg in logischen, aufeinanderfolgenden Ketten generiert werden. Das erkennt man zum Beispiel daran, ob die **Bewegungen** von Personen oder Objekten im Bild logisch und **fließend** verlaufen. Oder ob die **Lichtverhältnisse** konsistent und **realistisch** bleiben. Doch es ist davon auszugehen, dass solche Probleme zeitnah behoben sein werden.

Text-to-Video-Generator
Links: Durch neue Video-Generatoren wie „Sora“ (OpenAI) wird der Fotorealismus und die Qualität der Deepfakes immer besser. *Quelle: Screenshot, YouTube (2025)*
Rechts: Ein Nutzer teilt auf der Plattform X ein Deepfake-Video, um die Stärken des Modells „Gen-3 Alpha“ (Runway) zu demonstrieren. *Quelle: Screenshot, X (2025)*

Die Stimme eines jeden Menschen gilt, genau wie sein Fingerabdruck oder die Iris, als einzigartig. Ein Grund, warum viele Menschen solche biometrischen Merkmale als Methode zur „sicheren" Authentifizierung nutzen. Der Versuch, Stimmen mithilfe von Software künstlich zu erstellen oder zu klonen, galt lange Zeit als wenig erfolgversprechend. Softwarebasierte Stimmen klangen meist unnatürlich und „roboterhaft". Mittlerweile haben sich aber die Resultate bei der **Stimmsynthese** durch Deepfake-Technologien **erheblich verbessert**. Eine 2024 erschienene Studie der Universität Zürich hat gezeigt, dass KI-generierte Stimmklone das Potenzial haben, **von Menschen als „echt" wahrgenommen zu werden**.[7]

Um Audio-Deepfakes erstellen zu können, müssen in einem ersten Schritt, genau wie bei text- und bildbasierten Fakes, möglichst **viele Trainingsdaten** in hoher Qualität vorliegen. So „lernt" das Modell wie zum Beispiel die Zielstimme klingen soll. Dank unzähliger Beiträge auf YouTube und in den sozialen Netzwerken ist es heute relativ einfach, an solche Audiodaten für das Training der Modelle zu gelangen. Nach Abschluss des Trainings erhält das Modell dann die Inhalte, die es akustisch generieren soll. Es ist aber auch möglich, **Audio-Fakes auf Basis von wenigen Sekunden und vortrainierten Modellen** zu generieren.

Die Stimmsynthese erfolgt entweder per **Texteingabe** („Text-to-Speech") oder durch das **Verändern** („Voice Conversion") oder **Nachahmen echter Stimmen** („Voice Cloning").

Mit KI lassen sich nicht nur **menschenähnliche Stimmen** in diversen Sprachen künstlich erstellen, sondern auch **nonverbale Kommunikation** (Lachen, Seufzen, Weinen), **Umgebungsgeräusche, Soundeffekte und Musik**.

1 Text-to-Speech

Bei dem „Text-to-Speech"-Verfahren (kurz: TTS) werden **digitale Texte in gesprochene Sprache umgewandelt**. Diese Stimmen klingen oft neutral bzw. generisch und können keiner echten Person zugeordnet werden. Sie werden zum Beispiel in Sprachassistenten und Navigationssystemen eingesetzt. Man kann ein TTS-Modell aber auch mit Audiodaten einer bestimmten Person trainieren, um synthetische Stimmen zu erzeugen, die deren individuelle Sprachmuster imitieren. So kann der Eindruck erweckt werden, dass diese bestimmte Person den Text gesagt hat.

2 Voice Conversion (Speech-to-Speech)

Bei dem „Voice Conversion"-Verfahren (kurz: VC) wird eine **bestehende Sprachaufnahme in eine andere Stimme umgewandelt**. Die Originalaufnahme klingt nach der Veränderung so, als ob sie von einer anderen Person gesprochen wurde. Der **Inhalt bleibt dabei unberührt**. So kann zum Beispiel während eines (Video-)Telefonats der Eindruck erweckt werden, mit einer bestimmten Person zu sprechen.

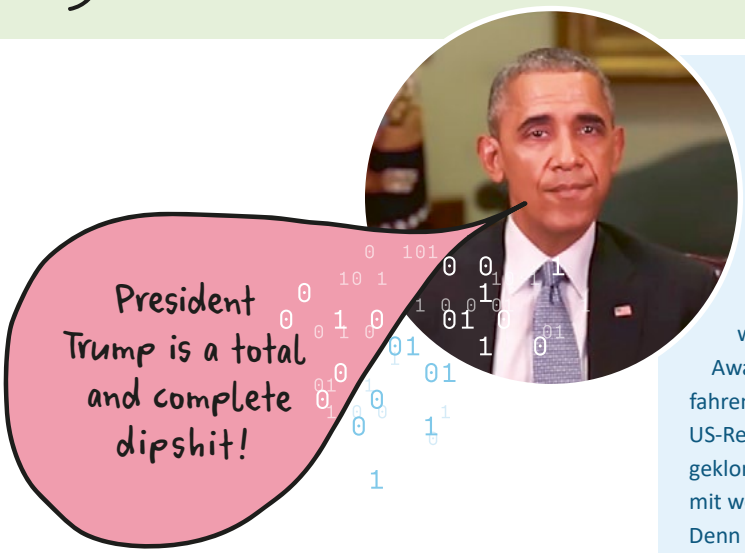
3 Voice Cloning

Bei dem Voice-Cloning-Verfahren wird eine **synthetische Stimmkopie einer bestimmten Person** erstellt. Hierfür werden Audiodateien der Zielperson benötigt, um das Modell darauf zu trainieren, wie die Person klingt und spricht. Anschließend können völlig neue, synthetische Sprachaufnahmen erzeugt werden, die jeden beliebigen Text wiedergeben und dabei wie diese Zielperson klingen.

STUDIE

Wie reagiert unser Gehirn auf Deepfake-Stimmen?

KI-generierte Stimmen kommen immer näher an die akustischen Signale echter Stimmen heran und können Menschen auch täuschen. Die Ergebnisse einer **Studie (2024) der Universität Zürich** zeigen auch, dass das **Gehirn auf Deepfake-Stimmen anders reagiert als auf natürliche Stimmen**. Das heißt: Rein kognitiv betrachtet verfügen Menschen über Mechanismen der Widerstandsfähigkeit gegenüber Stimm-Fakes. Auch wenn diese, den Ergebnissen zufolge, oft unter der Wahrnehmungsschwelle bleibt. Das zeigt, wie **wichtig** es ist, dass wir aktiv die eigenen Sinne schärfen und bewusst auf akustische Anomalien achten. Es gibt diverse Tools mit denen man sein **„kritisches Hören" gezielt trainieren** kann, zum Beispiel mit der Übung „Sharpen Your Senses!". Die Übung ist Teil des Projekts „Digger" und wurde vom Frauenhofer-Institut in Zusammenarbeit mit der Deutschen Welle und dem Athens Technology Center in Griechenland entwickelt.[8]



Audiobasierte Deepfakes

Ein bekanntes älteres Beispiel für Stimmkonvertierung ist das Deepfake-Video (2018), das den ehemaligen US-Präsidenten Barack Obama zeigt. Darin sieht und hört man, wie er seinen Nachfolger US-Präsident Donald Trump als „Vollidioten" bezeichnet. Doch das Video ist nicht echt, sondern wurde mithilfe von KI generiert. Es ist Teil einer Awareness-Kampagne von BuzzFeed, die auf die Gefahren durch Deepfakes hinweist. Tatsächlich spricht der US-Regisseur Jordan Peele diese Worte, allerdings mit der geklonten Stimme Obamas. Heute ließe sich der Deepfake mit weniger Aufwand und in höherer Qualität produzieren. Denn die Technologien in dem Bereich schreiten schnell voran. *Quelle: Screenshot, YouTube (2025)*

Warum ist nicht jeder Fake „deep"?

Aufgepasst! Nicht alles, was wie ein Deepfake klingt oder so aussieht, ist auch wirklich durch KI generiert oder verändert worden. Auch mit **herkömmlichen Programmen zur Bild- und Videobearbeitung** lassen sich Inhalte manipulieren. Wie beispielsweise das manuelle oder digitale Zusammenschneiden von Videoinhalten oder der Tausch von Gesichtern mithilfe von Photoshop. Solche Fälschungen sind technologisch betrachtet weitaus weniger anspruchsvoll als Deepfakes. Manchmal genügen sogar schon **einfache „Tricks"**, um gefälschte Medieninhalte zu erstellen. Indem man zum Beispiel Videoinhalte verlangsamt oder beschleunigt, gezielte Ausschnitte setzt oder Ähnlichkeiten zu bekannten Personen ausnutzt. Existierende Bild- und Videoaufnahmen lassen sich auch leicht in einen neuen Kontext setzen (Rekontextualisierung). Solche „simplen" Fakes werden daher auch als **„Cheap Fake"** oder „Shallow Fake" bezeichnet. Doch auch wenn solche Formen **visueller Desinformation** nicht „deep" sind, können sie einen großen Effekt haben.
Im Zuge des **US-Wahlkampfs 2020** verbreiteten sich einige Cheap Fakes in Social Media, die die Politikerin und US-Demokratin Nancy Pelosi verunglimpfen und als geistig ungeeignet für das politische Amt darstellen sollten. Ein Video, das Pelosi bei einem öffentlichen Auftritt zeigt, erweckt etwa den Eindruck, sie sei verwirrt. Einige kommentierten sogar, sie sei „betrunken". Die Originalaufnahme wurde manipuliert, indem die Geschwindigkeit des Videos verlangsamt wurde.

Cheap Fake

Trump teilt 2019 über die Plattform X dieses manipulierte Video, um die US-Politikerin Pelosi im Zuge des anstehenden Wahlkampfes (2020) herabzusetzen. *Quelle: Screenshot, X (2025)*

Ein weiteres Beispiel ist die Gegenüberstellung eines 2024 verbreiteten Cheap Fakes von (ehem.) US-Präsident Joe Biden mit dem Original-Video, das zeigt, wie mit einem „kleinen Trick" Inhalte gezielt aus dem Kontext gerissen werden. In dem manipulierten Video wird durch einen bewusst gesetzten Schnitt der Eindruck erweckt, Biden wolle sich auf einen nicht vorhandenen Stuhl setzen. Das ganze Video zeigt allerdings, dass es sehr wohl Stühle gibt.

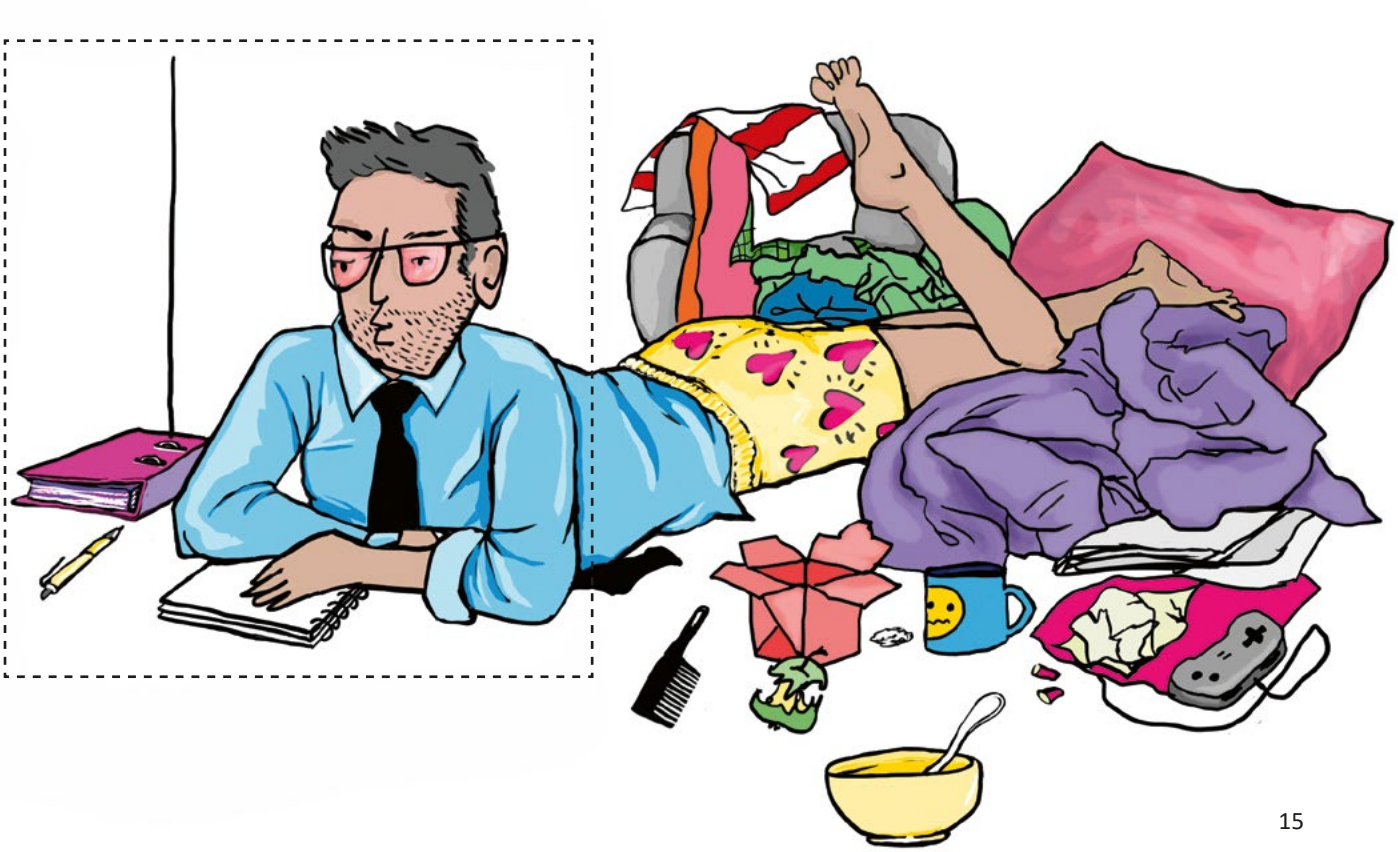
Demo: www.youtube.com/watch?v=pfDdfqFfEuk

Dadurch wirkt Pelosi träge und ihre Aussagen undeutlich. Bei einem anderen Video entsteht der Eindruck, dass Pelosi die ganze Zeit stammelt. Fakt ist aber, dass eine lange Rede der Politikerin gezielt zusammengeschnitten wurde. Der „billige" Fake wurde vom Trump-nahen US-Sender Fox News kommentiert und von Trump auf der Plattform X verbreitet.

Auch im **US-Wahlkampf 2024** wurden etliche „Cheap Fakes" über den damals amtierenden US-Präsidenten Joe Biden in Umlauf gebracht. Ziel war es, Biden als verwirrt und politisch handlungsunfähig erscheinen zu lassen, um die Bedenken potenzieller Wechsel-Wähler*innen weiter anzufeuern.

Die Illustration auf dieser Seite verdeutlicht, wie sich Gegebenheiten **durch gezieltes Framing rekontextualisieren lassen**. Denn manchmal reicht bereits ein gezielt gewählter Bildausschnitt aus, um die Wahrnehmung eines Medieninhalts zu manipulieren – ohne den Inhalt selbst verändern zu müssen.

Cheap Fakes zeigen, dass es keine ausgeklügelte Technologie braucht, um (politische) Desinformation zu betreiben. Auch mit einfachen Mitteln können Medieninhalte gezielt verändert und so aus dem Kontext gerissen werden.



Können Deepfakes sinnvoll sein?

Die Entwicklung und Anwendung von Deepfake-Technologien befinden sich, trotz der großen Fortschritte in jüngster Zeit, immer noch in den Kinderschuhen. Blickt man auf die mediale und öffentliche Debatte, so ist diese in erster Linie von den Sorgen über den missbräuchlichen Einsatz geprägt. Doch es gibt auch Branchen und Anwendungsfelder, in denen Deepfakes bereits relevant sind bzw. die damit etwas Nützliches verbinden.[9]

Unterhaltung

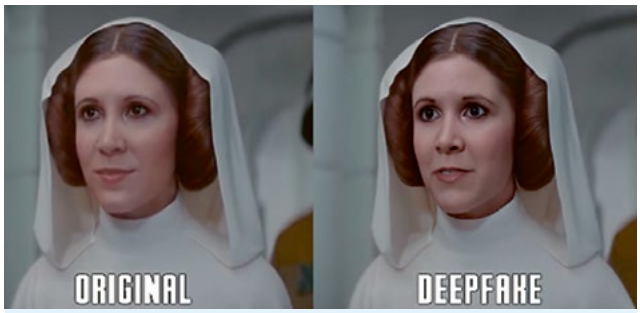
Filme

Vor einigen Jahren mussten noch aufwendige 3-D-Computergrafiken genutzt werden, um visuelle Tricks und Spezialeffekte umzusetzen. Mittlerweile kann vieles durch Deepfake-Technologien ergänzt bzw. sogar übernommen werden. Wobei die KI weitaus effizienter und kostengünstiger arbeitet und oft auch zu besseren Ergebnissen gelangt. **Große Hollywood-Studios** wie Walt Disney **setzen bereits Deep-Learning-basierte Verfahren ein** und treiben deren Forschung weiter voran.

Durch den Einsatz von Deepfake-Technologien ergeben sich für Film und Fernsehen vielseitige Möglichkeiten. Es lassen sich zum einen **Gesichter, Körper und Stimmen von Personen durch die anderer ersetzen**. Das heißt Stuntmänner bzw. -frauen können zum Beispiel per Face Swap wie die gewünschte Zielperson aussehen. Oder Schauspieler*innen können in beliebige Körper bzw. Figuren schlüpfen, ohne sich monatelang physisch auf eine Rolle vorzubereiten oder stundenlang in der Maske zu sitzen. Mithilfe von KI lassen sich zudem die **Gesichter von Schauspieler*innen verjüngen** („De-Aging"-Technologie genannt) und **verstorbene Filmstars zurück auf die Leinwand holen**.

Zum anderen können Filme **lippensynchron in andere Sprachen und Dialekte übersetzt oder mit Akzent versehen** werden, um ein vielfältigeres Publikum zu erreichen. Von lebenden Schauspieler*innen lassen sich auch **digitale Klone** erstellen. Diese können kommerziell genutzt werden, etwa in Kampagnen, Werbespots, Video-Games oder Hörbüchern, ohne dass die Person tatsächlich vor Ort sein und bestimmte Handlungen ausführen muss. Mit der Technologie lassen sich auch **fiktive, lebensechte Charaktere** erstellen, die zum Beispiel in der Rolle einer historischen Figur bestimmte Ereignisse nachstellen oder Interviews geben.

Manipulation von Filmen nicht neu
Schon seit Jahrzehnten ist es möglich, Videomaterial zu verändern. Wie zum Beispiel diese Szene aus dem Film „Forrest Gump" (1994). Die Filmemacher fügten Archivmaterial von John F. Kennedy digital in die Szene ein und manipulierten anschließend dessen Mundbewegung. Eingriffe wie diese, aber auch moderne Computerverfahren (CGI) zur Bildbearbeitung sind stets mit einem mühsamen und zeitaufwändigen Prozess verbunden. *Quelle: Screenshot, YouTube (2025)*



CGI-Verfahren vs. Deepfakes
Der Deepfaker „Shamook" zeigte auf seinem YouTube-Kanal zu was eine Person mithilfe von Deepfake-Technologie im Stande ist. Hier zu sehen am Film „Star Wars: Rogue One" (2016). Shamook erzielte mit seinen Deepfakes deutlich bessere Ergebnisse für eine verjüngte Prinzessin Leia (rechts) als Walt Disney (links). Seine Video-Fakes sind den mit Computer-generated Imagery (kurz: CGI) erstellten Varianten des Filmstudios, mit Blick auf Detailreichtum und Realismus, klar überlegen. Für Disney war die CGI-Arbeit zudem mit sehr großem technischen und personellen Aufwand verbunden. Shamooks arbeitet mittlerweile bei der Star-Wars-Produktionsfirma Lucasfilm, die zu Walt Disney gehört. *Quelle: Screenshot, YouTube (2025)*

Musik

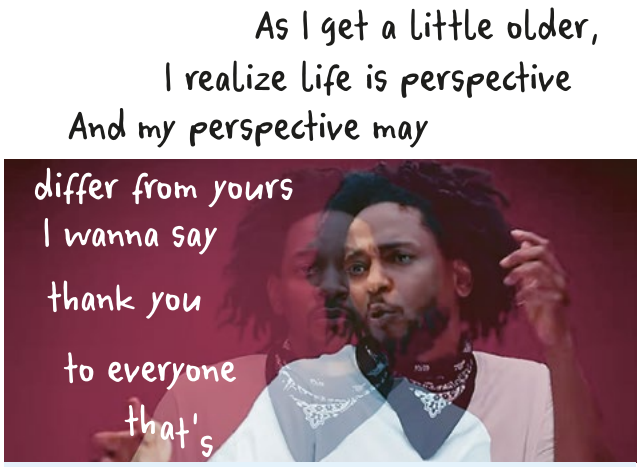
Auch für die Musik-Branche gibt es diverse Anwendungsmöglichkeiten von Deepfake-Technologien. Egal ob Songtexte, Stimmen, Kompositionen, Instrumente oder Soundeffekte – das alles und vieles mehr, kann mithilfe von KI erstellt bzw. nachgeahmt werden. So lassen sich **einzelne musikalische Komponenten oder auch ganze Musikstücke generieren**. Musiker*innen können solche Möglichkeiten auch als Hilfsmittel nutzen, um Ideen für neue Songs als **akustisches Memo** festzuhalten. Lieder lassen sich **in andere Sprachen übersetzen** und vorhandene Songs mit Blick auf **Klang, Aussprache und Betonung optimieren**. Zudem bieten Deepfakes für die Erstellung von **Musikvideos**, neue kreative Wege sich auszudrücken. So hat der Hip-Hop-Künstler Kendrick Lamar in seinem Musikvideo „The Heart Part 5“ (2022) Face Swaps als kreatives Stilmittel genutzt, um in verschiedene Rollen zu schlüpfen und dem Text so eine weitere Bedeutungsebene zu verleihen. Künstlerinnen könnten einen **digitalen Klon** von sich erstellen, der sie bei der Online-Interaktion mit Fans und Journalist*innen unterstützt – so wie es zum Beispiel Sängerin FKA Twigs getan hat. In der Rolle eines **kreativen Sparring-Partners** können Deepfake-Technologien Musiker*innen also die Möglichkeit bieten, kreativ und innovativ zu sein.

Gaming

Deepfakes können die Spielerfahrung im Gaming-Bereich positiv beeinflussen, indem sie **virtuelle Charaktere und Spielumgebungen realistischer, natürlicher und interaktiver** gestalten. Sind Mimik, Sprachausgabe sowie Lippensynchronisation verbessert, wirken Gesichtsausdrücke und Reaktionen der Charaktere lebensnaher und authentischer. Eine verbesserte Darstellung von Emotionen sowie dynamische und natürliche Dialoge zwischen den Spieler*innen können für eine tiefere Interaktion zwischen den Spielfiguren und eine stärkere Bindung in das Spiel sorgen. Darüber hinaus tragen natürliche Körperbewegungen sowie die Möglichkeit, einen digitalen realistischen Avatar von sich zu erstellen, zu mehr Authentizität und einer stärkeren Immersion der Spieler*innen in das Spielgeschehen bei.

Fan Art

Deepfakes werden in der Fan Community oft auf kreative Weise genutzt. Zum Beispiel um **Charaktere** in beliebten Filmen **alternativ darzustellen** oder um **alternative Szenen** zu erstellen, die in der ursprünglichen Geschichte gar nicht vorkommen. Sehr beliebt ist auch, Deepfakes als **Mittel für Humor und Satire** zu nutzen.



Kreativer Einsatz in Musikvideos
Kendrick Lamar rappt in seinem Song „The Heart Part 5“ über moralische Fragen, die er anhand von prominenten, öffentlich kontrovers diskutierten Afroamerikanern unter die Lupe nimmt. Dabei schlüpft Lamar mithilfe von Deepfake-Technologie auch visuell in die Rolle der insgesamt sechs Personen, über die er singt. Es wirkt, als sängen die mithilfe von KI-generierten Personen selbst diese Zeilen, um ihre Sichtweise zu spiegeln. *Quelle: Screenshot, YouTube, bearbeitet von klicksafe (2025)*

Deepfake-Memes
Links: Joe Biden und Donald Trump als „gute Freunde“ auf gemeinsamer Urlaubstour. *Quelle: Screenshot, YouTube (2025). Demo: www.youtube.com/shorts/K2z_Q07hlog*
Rechts: Bekannte Persönlichkeiten wie Arnold Schwarzenegger sind beliebtes Ziel humorvoller, satirischer oder ironischer Deepfakes, die viral gehen. Schwarzenegger als Frodo Beutlin in Herr der Ringe. *Quelle: Screenshot, YouTube (2025).*
Demo: youtu.be/jgETccDDp8E?feature=shared

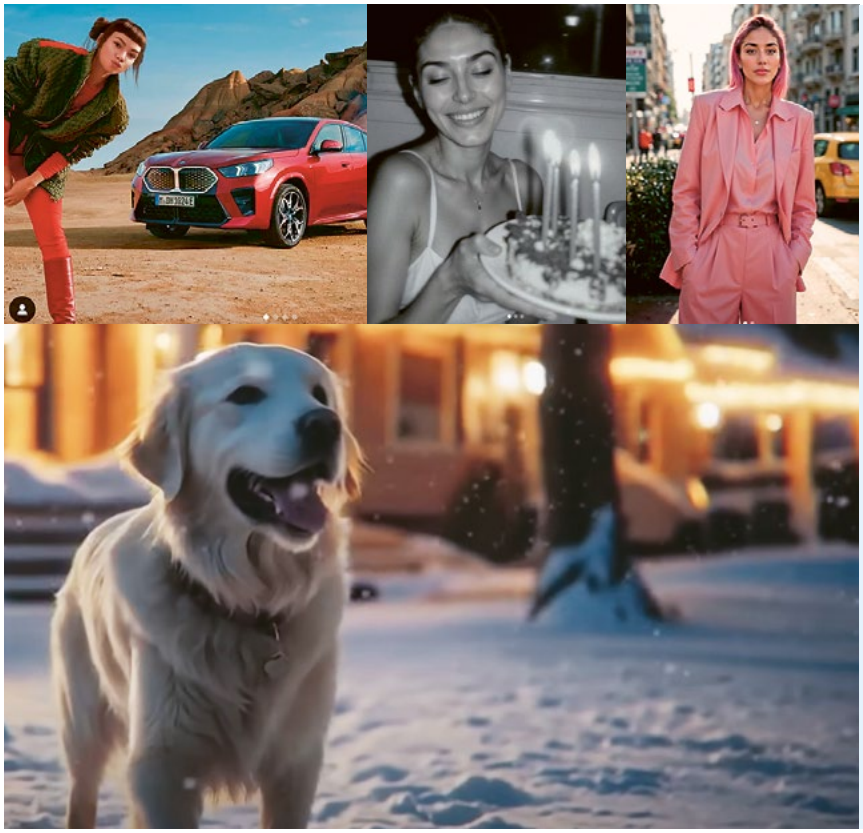
Kommerzieller Nutzen

Durch die technologischen Fortschritte in den letzten Jahren haben sich die Fähigkeiten **virtueller Charaktere** (KI-Avatare) enorm weiterentwickelt. Vor allem in Bezug auf:

- **Kosten:** Die Kosten zur Erstellung von Avataren sind gesunken. Damit steigt die Zielgruppe, die sich den Einsatz solcher Charaktere leisten können.
- **Qualität:** Das Aussehen, Verhalten, die Sprache und Emotionen digitaler Avatare wirken zunehmend authentisch, was sie insgesamt realistischer erscheinen lässt.
- **Anpassungsfähigkeit:** Virtuelle Charaktere können noch stärker personalisiert werden, sprich noch besser die Wünsche und Bedürfnisse der Nutzenden berücksichtigen.
- **Interaktivität:** Mit Avataren sind zunehmend dynamische, fast menschenähnliche Gespräche möglich.
- **Immersion:** Durch all diese Fortschritte können virtuelle Charaktere zunehmend als „real" empfunden werden.

Damit ergeben sich ganz neue Möglichkeiten für den Einsatz von hyperrealistischen Avataren. Im **Journalismus** können sie zum Beispiel zur Moderation von Nachrichten eingesetzt werden (z. B. Wetter, Sport, Verkehr). Dies ist vereinzelt sogar bereits der Fall. Die Technologie kann auch dabei helfen, die Interaktion zwischen **digitalen Assistenten** und den Nutzenden zu verbessern.

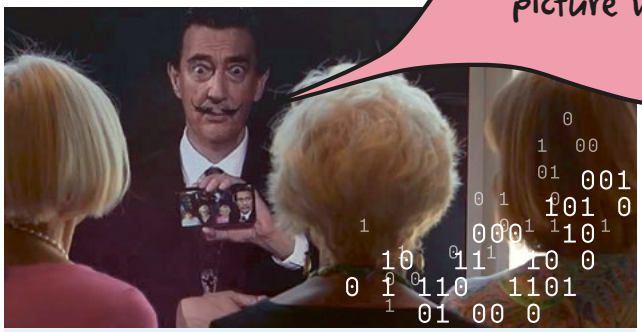
Vor allem im **Online-Handel** gilt der Einsatz von KI-Avataren als ein wichtiger Zukunftsmarkt. Zum einen, um den **Kunden-Service** zu verbessern. Ein Vorhaben, das bei Modeketten wie H&M und Zalando aktuell in der Pilotphase steckt, sind sogenannte **virtuelle Umkleidekabinen**. Kund*innen können mit dem Smartphone einen 3D-Ganzkörper-Avatar von sich erstellen und mit diesem virtuell Produkte ausprobieren. So sollen auch Fehlkäufe vermieden und die hohe Zahl an Retouren gesenkt werden.
Zum anderen spielen KI-Avatare in der **Werbung** eine zunehmend große Rolle. Das zeigt sich zum Beispiel daran, dass ganze **Werbespots** KI-generiert sind oder auch am Phänomen der **KI-Influencer*innen**. Das sind virtuelle Charaktere mit einem Social-Media-Auftritt, die mithilfe von Deepfake-Technologien erstellt und in der Regel für kommerzielle Zwecke eingesetzt werden. Bekannte Charaktere wie Lil Miquela, Aitana Lopez, Shudu oder Imma interagieren mit ihren Fans, posten Inhalte oder gehen Kooperationen mit großen Marken ein, genau wie echte Influencer*innen. Modehäuser wie Mango oder Etro haben 2024 gezeigt, wie sich **KI-Models** gezielt **für Werbekampagnen** einsetzen lassen. Zum einen, um innovative Bildinhalte umzusetzen, die sich noch mehr von materiellen Grenzen lösen. Zum anderen aber auch, um Kosten in der Wertschöpfungskette einzusparen.



Deepfakes in der Werbung
Links oben: KI-Influencerin Lil Miquela wirbt u. a. für große Marken wie Prada, Calvin Klein, Diesel oder BMW. *Quelle: Screenshots, Instagram (2025)*
Mitte und rechts oben: KI-Influencerin Aitana Lopez wirkt auf vielen Bildern wie eine echte Person. Die Accounts beider KI-Influencerinnen sind aber mit dem Hinweis gekennzeichnet, dass sie KI-generiert sind. *Quelle: Screenshots, Instagram (2025)*
Unten: Wie jedes Jahr zeigte Coca Cola auch 2024 einen Weihnachtswerbespot. Das Besondere war diesmal, dass der Spot vollständig mit künstlicher Intelligenz erstellt wurde. *Quelle: Screenshot, YouTube (2025).*
Demo: www.youtu.be/4RSTupbfGog?feature=shared

Bildung

Im Bildungsbereich können Deepfake-Technologien ebenfalls einen nützlichen Beitrag leisten. So lassen sie sich über alle Bildungswege hinweg dazu einsetzen, um **Bildungsangebote ansprechender, anschaulicher und immersiver** zu gestalten. Zum Beispiel, indem eine historische Persönlichkeit oder ein Zeitzeuge mithilfe eines Chatbots oder KI-generierten Avatars virtuell zum Leben erweckt wird. Schüler*innen können mit dieser Figur in Form eines fiktiven Gesprächs interagieren, um mehr zu der Person, ihrem Hintergrund und der Zeit zu erfahren, in der sie gelebt hat.[10]



Deepfakes in Museen
Der Künstler Salvador Dalí als digitaler Avatar in der Ausstellung „Dalí Lives“. Der Avatar wurde mithilfe von KI generiert und interagiert mit seinem Publikum. Jedes Gespräch verläuft ein wenig anders, was die Erfahrung noch persönlicher macht. *Quelle: Screenshot, YouTube (2025)*

Deepfakes für Aufklärungsarbeit
Die globale Kampagne „Malaria Must Die“ wurde von „Malaria No More UK“ in Zusammenarbeit mit dem ehemaligen Fußballer David Beckham und dem KI-Unternehmen Synthesia umgesetzt. In dem Video werden Beckhams Worte mithilfe der Synchronisationstechnologie von Synthesia nahtlos in neun Sprachen übersetzt, inklusive Lippensynchronisation. *Quelle: Screenshot, YouTube (2025)*

Auch **Lernorte außerhalb des formalen Schulsystems** wie Museen oder Bibliotheken können von solchen Ansätzen der Wissensvermittlung profitieren. Ein Beispiel ist die Ausstellung „Dalí Lives“, die den verstorbenen Künstler anhand von Archivmaterial und KI zum Leben erweckt. Der Avatar interagiert mit den Besucher*innen und erzählt Geschichten über sein Leben als Künstler. Am Ende fragt der virtuelle Dalí sogar, ob er mit ihnen ein Selfie machen darf, was sie sich mit dem Smartphone dann herunterladen können. **Deepfakes können also dazu beitragen, Geschichten zu erzählen und Kultur zu bewahren**.

Des Weiteren können Deepfakes in der **medienpädagogischen Arbeit** gezielt dazu eingesetzt werden, um junge Menschen für die Technologien sowie die damit verbundenen Risiken zu sensibilisieren. Dies kann auch über Online-Kampagnen erfolgen, die eine breitere Öffentlichkeit erreichen. Ein bekanntes Beispiel wäre auch hier wieder die manipulierte Rede des ehemaligen US-Präsidenten Barack Obama. Deepfakes können auch für **Aufklärungsarbeit und Öffentlichkeitskampagnen** eingesetzt werden. Indem Botschaften lippensynchron übersetzt und zielgruppenspezifisch angepasst werden können, lässt sich ein diverseres Publikum ansprechen und eine höhere Reichweite erzielen.

WORKSHOPS

Mit medienpädagogischen Workshops auf Deepfakes aufmerksam machen

techagogics bietet kostenlose immersive Bildungsworkshops an mit pädagogischen Ansätzen für Kinder der Klassenstufe fünf bis zehn. In dem Workshop „Deepfake Detective“ lernen sie, wie sie Deepfakes erkennen können. Weitere Infos unter:
→ **www.deepfake-detective.de**

Empowerment

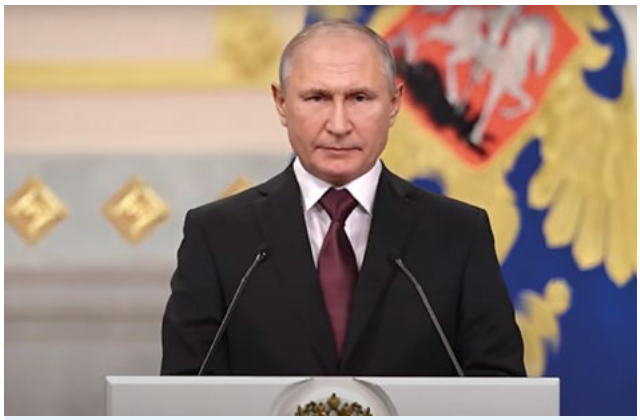
In bestimmten Kontexten könnten Deepfakes auch dazu beitragen, dass Menschen zu mehr Selbstbestimmung und Handlungsfähigkeit gelangen. Sei es durch das **Senken von Barrieren**, indem Deepfakes beispielsweise die lippensynchrone Übersetzung in viele Sprachen ermöglicht oder ein Avatar Menschen mit Hörbeeinträchtigung in Gebärdensprache unterstützt. Oder durch **unterstützende Systeme** wie einen personalisierten interaktiven Avatar, der ältere oder beeinträchtigte Menschen im Alltag unterstützt. Auch ein **Schutz von Identitäten** ist möglich, zum Beispiel durch die Anonymisierung von Personen in Zeugenschutzbefragungen oder journalistischen Interviews. Deepfake-Tools bieten auch **neue Ausdrucksmöglichkeiten**, die theoretisch jede Person nutzen kann, um etwa kreativ oder politisch aktiv zu sein. Ohne, dass spezielle Fachkenntnisse notwendig sind, wie etwa Wissen über die Produktion von Medieninhalten.

Gesundheit und Wohlbefinden

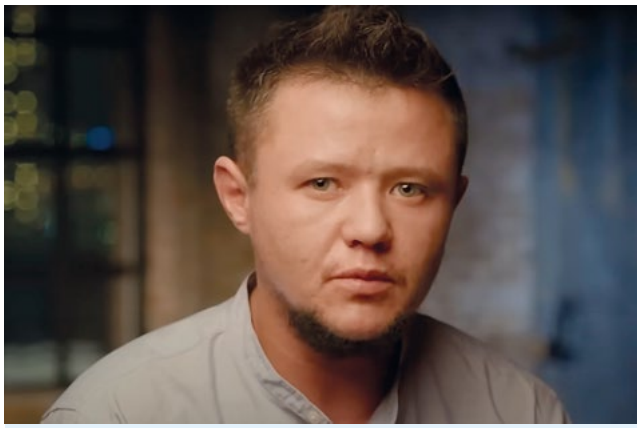
KI-Avatare können auch im Gesundheitswesen, etwa in der **Psychotherapie** eingesetzt werden. So zeigen Forscher*innen, dass sich Deepfake-Technologie positiv in der Therapie von psychischen Erkrankungen einsetzen lassen. Zum Beispiel bei Trauer, Depressionen, Traumata oder posttraumatischer Belastungsstörung.[11] Mithilfe von KI-Avataren können zum Beispiel Opfer von Mobbing in einem geschützten Rahmen mit den Täter*innen konfrontiert werden. Oder Menschen mithilfe von verjüngten Deepfake-Versionen von sich selbst, immersive Innere-Kind-Arbeit machen. Deepfakes können auch dazu eingesetzt werden, um mit dem Verlust geliebter Menschen umzugehen, indem sie digital wieder „zum Leben" erweckt werden. KI-Avatare stehen auch in der Diskussion, als Werkzeug zur Kommunikation eingesetzt zu werden, um **Einsamkeit zu überwinden**.

SITESTEP

Auch hinter manchen Gesichtsfiltern auf dem Handy oder in Social Media stecken Deepfake-Technologien. Durch solche Filter können sich Nutzende visuell verändern, indem sie sich zum Beispiel in ein Tier oder eine Cartoon-Figur verwandeln. Tipps zum Umgang mit Filtern gibt es hier:
→ **www.klicksafe.de/mnsn00**
→ **https://t1p.de/mfdga (elternguide.online)**



Deepfakes im Einsatz für politischen Aktivismus
Einen umstrittenen Ansatz wählte die Organisation „Represent US" in ihrer Anti-Korruptionskampagne (2020), um die Bürger*innen in den USA für die bevorstehenden Wahlen zu sensibilisieren. Die gefakten Botschaften stammen vermeintlich aus dem Mund von Wladimir Putin und in einem anderen Clip angeblich von Kim Jong-Un. Am Ende jedes Clips erscheint der Satz: „Democracy lives or dies with you.", um auf die Bedeutung von Wahlen in Demokratien aufmerksam zu machen. *Quelle: Screenshot, YouTube, (2025).*
Demo: www.youtube.com/watch?v=sbFHhpYU15w



Deepfakes zum Schutz von Personen
Der Journalist und Filmemacher David Frances nutzte zum Beispiel Deepfakes in seinem Dokumentarfilm „Welcome to Chechnya", der die Verfolgung von LGBTQ+-Menschen in Tschetschenien thematisiert. Durch Face-Swap-Verfahren konnten die Porträtierten anonym bleiben, während ihre Emotionen im Gesicht erhalten blieben.
Quelle: Screenshot, YouTube, (2025).
Demo: youtu.be/_2KMm49B6pE?feature=shared

Können Deepfakes gefährlich sein?

In der öffentlichen Debatte wird mit Deepfakes vor allem die Sorge verbunden, dass diese eingesetzt werden können, um Menschen gezielt zu beeinflussen oder ihnen sogar zu schaden. Im Folgenden werden die verschiedenen Anwendungsfelder skizziert, in denen Deepfakes bereits jetzt missbräuchlich zum Einsatz kommen.[12]

Cybermobbing

Deepfakes sind eine neuere Variante, um Cybermobbing zu begehen. Dabei werden einzelne Individuen mit KI-generierten Inhalten **gezielt** und über einen längeren Zeitraum **online beleidigt, bloßgestellt, belästigt oder sogar bedroht**. Dies kann theoretisch über alle Plattformen geschehen, die eine Interaktion sowie das Hochladen von Medieninhalten erlauben – angefangen von Online-Game-Plattformen, über soziale Netzwerke und Messenger, bis hin zu Videoportalen. In der Regel werden Betroffene von Personen schikaniert, die sie auch persönlich kennen.
Mobbing-Attacken durch Deepfakes lassen sich durch den leichten Zugang zu diversen Deepfake-Tools im Netz recht **einfach umsetzen**. Alles, was es dazu braucht, sind Bilder, Videos oder Audioinhalte der Betroffenen. Diese sind zum Beispiel über soziale Netzwerke, Messenger-Dienste oder heimliche Ton- und Bildaufnahmen zugänglich. Über KI-Tools können solche Aufnahmen verändert, aus dem Kontext gerissen oder mit anderen Medieninhalten kombiniert werden.

MEHR INFOS

Unter diesem Link finden Sie weitere Informationen, Beratungsangebote für pädagogische Fachkräfte, Eltern und Jugendliche sowie Sofortmaßnahmen, um von Cybermobbing betroffenen Kindern zu helfen:
→ **www.klicksafe.de/cybermobbing**

MEDIALE FALLBEISPIELE

Cybermobbing mit Deepnudes

In Spanien wurde 2023 ein Fall medial bekannt, indem über 20 minderjährige Schülerinnen Opfer von KI-generierten Nacktbildern wurden, die Mitschüler und Bekannte im Netz verbreitet hatten. Auch in den USA gab es 2024 einen Fall, bei dem ein Schüler KI-generierte Nacktbilder von einigen Mädchen erstellt und in Umlauf gebracht hatte. In den USA wurde 2021 sogar eine Mutter wegen Verleumdung durch Deepfakes angeklagt. Ihr wurde vorgeworfen, kompromittierende Deepfake-Bilder und Videos von den Cheerleader-Konkurrentinnen ihrer Tochter erstellt und verbreitet zu haben. Sie soll damit das Ziel verfolgt haben, dass diese aus dem Team ausgeschlossen werden.

Mit dem Ergebnis, dass man Betroffene bei scheinbaren Handlungen sieht, die faktisch nie passiert sind oder Worte aus ihrem Mund hört, die sie nie gesagt haben. Solche Deepfakes, egal ob amateurhaft oder täuschend echt umgesetzt, haben das Ziel, den Ruf der Person zu schädigen.

Hass und Hetze

Noch viel häufiger werden Deepfakes dazu genutzt, um Hass und Hetze gegen bestimmte Menschengruppen im Internet zu verbreiten. Hassrede richtet sich gegen Individuen oder Gruppen, die **aufgrund bestimmter Merkmale diskriminiert und abgewertet** werden. Etwa aufgrund von Herkunft, Hautfarbe, Geschlecht, sexueller Orientierung, Religion oder politischer Haltung. Typische Muster von Hassrede sind zum Beispiel Verallgemeinerungen („Alle X sind kriminell.“), das Verwenden von Stereotypen und Vorurteilen durch bestimmte Begriffe und Metaphern („Asylantenflut“) und die bewusste Verbreitung von Lügen (Desinformation). Typisch ist auch,

KI-ASSISTENT „GROK" AUF X – WENIG ETHISCHE LEITPLANKEN

Grok ist ein KI-Assistent der Firma xAI, die von **Elon Musk** gegründet wurde. Musk tritt damit in Konkurrenz zu OpenAI und ChatGPT, mit dem selbst erklärten Ziel, eine ungefilterte und „maximal wahrheitssuchende" KI zu etablieren. „Grok" ist eine Wortneuschöpfung und bedeutet im amerikanischen Slang „kapieren". Der Chatbot startete zunächst als reine Konversations-KI auf der Plattform X und wurde Ende 2024 um den **Bildgenerator „Aurora"** erweitert. Anfangs gab es für den Generator kaum Leitplanken, was es leicht machte, gewalttätige, sexuelle oder irreführende Bilder zu generieren. Es gab öffentlich viel Kritik, weshalb nachträglich gewisse Sicherheitsschranken in das Tool eingebaut wurden. Aktuell ist es zum Beispiel nicht mehr möglich, Personen mit Waffen oder beim Drogenkonsum zu generieren. Anders als bei vielen konkurrierenden KI-Tools lassen sich bei Grok Menschen generieren, die es wirklich gibt – zum Beispiel Personen aus der Öffentlichkeit. Die Bilder haben zum Teil **fotorealistische Qualität**. Bemerkenswert ist, dass für das Modell-Training und die Feinabstimmung des Chatbots auch auf **Nutzerdaten der Plattform X** zurückgegriffen wird. Es sei denn, man deaktiviert die Option für die Verarbeitung der eigenen Daten in den Systemeinstellungen. Dort heißt es: „X may share with xAI your public X data as well as your user interactions, inputs and results with Grok on X to train and fine-tune Grok and other generative AI models." Der Chatbot kann auch in Echtzeit auf öffentliche Beiträge auf X zurückgreifen. Ziel soll dabei sein, aktuelle „Informationen" miteinzubeziehen, wenn Fragen der Nutzenden beantwortet werden. Die Plattform X (ehemals Twitter) ist seit der Übernahme durch Musk äußerst umstritten, da dort Hass, Lügen und rechtspopulistische Inhalte gezielt Raum erhalten. Musk selbst sowie US-Präsident Trump tragen dort zur Verbreitung solcher Inhalte massiv bei.

Mit Deepfakes gezielt Hassrede verbreiten

Der KI-Assistent „Grok" wird auch zur Verbreitung von diskriminierenden Inhalten genutzt. Es gibt zum Beispiel ein Video, das die Präsidentschaftskandidatin Kamala Harris mit nacktem Babybauch zeigt, augenscheinlich schwanger von Trump. Ein anderes Bild (unten) zeigt Trump, wie er Haustiere rettet, verfolgt von zwei Haitianern. Bilder wie dieses gingen viral, nachdem Trump öffentlich behauptet hatte, dass haitianische Einwanderer (bestimmte Menschengruppe) die Haustiere von Amerikaner*innen essen und eine Bedrohung darstellen. Beide Beispiele wurden mithilfe von KI generiert und enthalten typische Muster von Hassrede wie Humor, Ironie, Sarkasmus, Desinformation, Stereotype, Vorurteile und Verallgemeinerung. *Quelle: Screenshot, X (2025)*

Mit Deepfakes gezielt Frauen verleumden

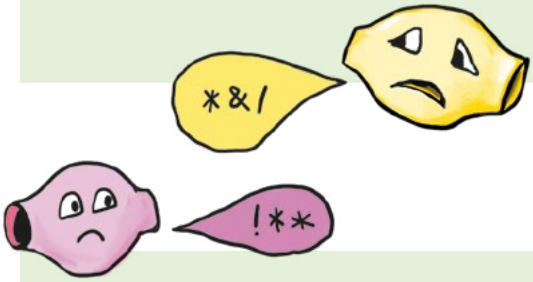
Als Joe Biden seinen Rücktritt aus dem US-Wahlkampf 2024 erklärte, sprach er sich für die ehem. US-Vizepräsidentin Kamala Harris als seine Nachfolgerin aus. Daraufhin gingen im Wahlkampf viele Deepfakes viral, die Harris als Präsidentschaftskandidatin degradierten. Ein mithilfe von KI erstelltes Video zeigt sie etwa als Liebespartnerin von Donald Trump – händchenhaltend, sich küssend und mit nacktem Babybauch. Es gab auch Deepfakes, die Harris nur in einem Bikini bekleidet zeigten, eng umschlungen mit dem verurteilten und verstorbenen Sexualstraftäter Jeffrey Epstein. *Quelle: Screenshot, Instagram (2025)*

Verschwörungstheorien zu verbreiten („9/11 war ein Inside Job“), menschenverachtende Aussagen mit Humor, Ironie und Sarkasmus zu tarnen oder Gewalttaten zu befürworten bzw. sogar dazu aufzurufen („Alle an die Wand stellen.“). **Durch Deepfake-Technologien lassen sich diskriminierende Botschaften leicht ins Audiovisuelle übertragen**. Ganz gleich, ob sie auf den ersten Blick klar erkennbar sind oder durch eine subtile Bildsprache (z. B. Codes) vermittelt werden. Deepfakes eröffnen einen neuen Spielraum, in dem hasserfüllte Botschaften schnell und in Masse erzeugt werden können.

STUDIE

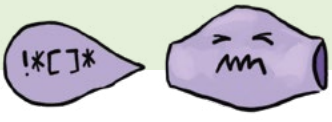
Hass im Netz bedroht den demokratischen Diskurs

Jeden Tag werden Menschen im Netz beleidigt, belästigt und bedroht. Viele ziehen sich bereits zurück und äußern ihre politische Meinung dort seltener. Das gefährdet Meinungsvielfalt und Demokratie. Die Studie „Lauter Hass – leiser Rückzug“ analysiert die Erfahrungen deutscher Internetnutzer*innen mit Hass im Netz und liefert hierzu aktuelle Zahlen und Fakten.
→ **www.toneshift.org/lauter-hass-leiser-rueckzug**



MEHR INFOS

Unter diesem Link finden Sie weitere Informationen für pädagogische Fachkräfte, Hilfs- und Beratungsangebote für Jugendliche sowie ein Quiz zum Thema Hate Speech:
→ **www.klicksafe.de/hate-speech**



Sexuelle Gewalt und Frauenhass

Viele Menschen haben selbst schon Online-Hass erlebt. Doch nicht alle sind davon in gleicher Weise betroffen. Studien zeigen, dass **besonders junge Frauen Zielscheibe von Hass im Netz** sind. In dem Kontext spielen auch **Deepfakes eine zunehmend problematische Rolle**.

- Bei sogenannten **Deepnudes** handelt es sich um KI-generierte Nacktbilder, die mithilfe von Deep Learning, einer Methode des maschinellen Lernens erstellt werden. Online gibt es zahlreiche Webseiten und Apps, die aus normalen, nicht sexualisierten Fotos realistisch wirkende Nacktbilder machen können. Hierzu erfassen die Programme den Körper der betroffenen Person und erstellen ein neues, entkleidetes Bild von dieser. Sie werden gezielt zum Zweck der virtuellen Entkleidung programmiert und stehen theoretisch jeder Person kostenlos bzw. zu geringen Kosten zur Verfügung. Auch Jugendliche können problemlos auf solche Dienste zugreifen und kompromittierende, gefälschte Bildinhalte von anderen erstellen.
- Mit Deepfake-Technologien lässt sich auch **Deepfake-Pornografie** erstellen. In der Regel basieren solche Fakes auf dem Austausch von Gesichtern (Face Swapping). Dabei wird das Gesicht einer realen Person nahtlos in ein bereits bestehendes pornografisches Video hinein montiert. Pornografische Deepfakes machen einen Großteil aller Deepfakes aus, die im Netz kursieren. Bei den Betroffenen handelt es sich in der Regel um (junge) Frauen. Bislang haben vor allem Frauen, die in der (politischen) Öffentlichkeit stehen, ein höheres Risiko, in Deepfake-Pornos abgebildet zu werden. Doch das Phänomen betrifft auch zunehmend Privatpersonen.
- Deepfakes verstärken auch die Gefahr von **Sextortion**. Das ist eine Form der digitalen Erpressung, bei der die Täter*innen die Betroffenen mit intimen Bildern oder Videos unter Druck setzen. Zum Beispiel um an Geld oder (weitere) Intimaufnahmen zu gelangen. Da sich mithilfe von KI vermeintlich intime Fotos oder Videos erstellen lassen, sind für solche Erpressungen nicht mehr echte Intimaufnahmen der betroffenen Person notwendig.

Bildbasierte sexualisierte Gewalt für Betroffene hochgradig belastend

Obgleich Deepnudes und Deepfake-Pornos „nur“ Fälschungen sind, handelt es sich bei beiden Formen um sexuelle Gewalt, die für Betroffene äußerst belastend sein können. Denn sie sind einer Situation ausgeliefert, in der jemand sie **gezielt bloßstellt und erniedrigt**. Werden die Bildinhalte verbreitet, kann das für Betroffene extrem **rufschädigend** sein oder auch **soziale Ausgrenzung** zur Folge haben. Viele leben daher in stetiger Angst, dass noch mehr Menschen die Inhalte sehen könnten. Wenn Deepfakes „gut gemacht“ sind, ist es für Betroffene zudem schwer, andere davon zu überzeugen,

LINKTIPP

Aufklärungsvideo von JUUUPORT

Das Video mit dem Titel „Täuschend echt! Was Du über Deepfakes wissen solltest“ sensibilisiert Jugendliche niedrigschwellig für die Risiken von KI-generierten Fälschungen im Internet wie die Erpressung mit Nacktaufnahmen und Cybermobbing. Zudem zeigt es Hilfsmöglichkeiten auf. Das Video wurde mit Unterstützung der Bayerischen Landeszentrale für neue Medien (kurz: BLM) produziert.
→ www.youtube.com/watch?v=WERLcclobgU

dass die Bilder bzw. Videos nicht echt sind. Sind dann auch noch private Daten mit den Fakes geteilt worden, wie Klarnamen und Wohnanschrift (sogenanntes Doxing) kann das für viele ein **Verlust des Sicherheitsgefühls** zur Folge haben. Aus Scham und Angst vor den Reaktionen der Eltern kann es gerade für junge Menschen eine große Hürde sein, sich als Betroffene jemandem anzuvertrauen.

Erstellen pornografischer Deepfakes über unterschiedliche Wege möglich

Pornografische Bilder und Videos können online über verschiedene Wege erstellt werden. Es gibt zum einen zahlreiche Angebote, die dafür werben, pornografische Inhalte zu erstellen. Online finden sich auch viele **Tutorials**, die mit einer Schritt-für-Schritt-Anleitung Interessierten dabei helfen, solche Angebote zu nutzen. Über **Serviceportale** können Bilder auch hochgeladen und Deepfakes zu geringen Kosten gekauft werden. Zum anderen bieten einzelne **Deepfaker** ihre kostenpflichtigen Dienstleistungen in Online-Foren und Marktplätzen an. Dieses breite Angebot hat zur Folge, dass theoretisch jede Person solche Inhalte erstellen (lassen) kann.

Spaß, Rache, Demokratieverzerrung – Motive hinter pornografischen Deepfakes divers

Die **Beweggründe**, die hinter der Erstellung von pornografischen Deepfakes stecken, können **vielfältig** sein. Angefangen von „Spaß“, sexueller Neugier oder Befriedigung sexueller Fantasien, bis hin zu Cybermobbing, Rache an der Ex-Partnerin oder Erpressung. Dass Deepfakes gezielt gegen Frauen eingesetzt werden, kann aber auch mit einem Hass auf bzw. einer generellen **Abwertung von Frauen und Mädchen** („Misogynie“) zusammenhängen. Geschlechtsspezifische Gewalt ist nicht neu, erreicht aber **durch pornografische Deepfakes eine völlig neue Dimension**.

Kompromittierende Deepfakes zielen dabei oft auf die Existenz von **Frauen im öffentlichen Raum** ab. Das heißt, dass vor allem jene mit sexualisierten Inhalten gedemütigt und eingeschüchtert werden, die ihre Themen und Positionen im Netz sichtbar machen – besonders wenn sie feministisch sind. Egal ob es sich dabei um Politikerinnen, Wissenschaftlerinnen, Journalistinnen, Aktivistinnen, Influencerinnen oder Frauen aus der Zivilgesellschaft, Kunst- und Kulturszene handelt. **Das Ziel solcher Attacken ist, (junge) Frauen im Digitalen systematisch anzugreifen und so aus dem öffentlichen Diskurs zurückzudrängen.** Und tatsächlich: Aus Angst vor digitaler Gewalt neigen vor allem Frauen dazu, sich aus sozialen Netzwerken zurückzuziehen. Jede dritte Frau befürchtet sogar, dass von ihr im Netz gefälschte Nacktbilder oder intime Aufnahmen ohne Einwilligung veröffentlicht werden könnten (vgl. Hate Aid, 2021). Das heißt: Pornografische Deepfakes können weitreichende gesellschaftliche Folgen haben. Denn wenn Frauen aus Angst vor Betroffenheit und Selbstschutz ihr Onlineverhalten anpassen, einschränken oder sogar komplett einstellen, führt dies zu einer **geringeren Meinungsvielfalt im Netz** – auch „Silencing“ genannt (dt.: Verstummen). Das hat zur Folge, dass Frauen ihr **Recht auf freie Meinungsäußerung und digitale Teilhabe nicht voll wahrnehmen können**.

MEHR INFOS

Unter diesem Link finden Sie weitere Informationen, Tipps und Handlungsmöglichkeiten für Betroffene und ihre Unterstützer*innen sowie passende Hilfe- und Beratungsstellen:
→ www.klicksafe.de/sexualisierte-gewalt-durch-bilder

Missbrauch von Identitäten und Betrug

Dokumente fälschen, sich als eine andere Person ausgeben, Vertrauen missbrauchen, Passwörter „knacken" – all das ist nicht neu. Doch durch Deepfake-Technologien haben sich die Methoden der Erstellung sowie die Qualität von Fälschungen erheblich weiterentwickelt. Diese Möglichkeiten machen sich auch Kriminelle zu Nutze. Für **Privatpersonen** ergeben sich daraus **verschiedene Risiken**, wie:

- **Social Engineering-Angriffe:** „Social Engineering" beschreibt das psychologische Manipulieren von Personen, um an vertrauliche Daten zu gelangen. Dazu zählt auch „Phishing" – eine Taktik, bei der Personen über E-Mails, Kurznachrichten oder andere Kommunikationswege kontaktiert und zu einer bestimmten Handlung aufgefordert werden. Diese Nachrichten können in ihrer Aufmachung so wirken, als stammten sie aus einer vertrauenswürdigen Quelle. Ziel ist, das Vertrauen der Betroffenen zu gewinnen, um an sensible Daten zu gelangen (zum Beispiel an Passwörter, Bankdaten oder persönliche Informationen). Deepfakes machen solche Täuschungen noch glaubwürdiger. Gerade durch Text-Fakes lassen sich Stilistik und Struktur eines vertrauenswürdigen Formats bzw. Institution leicht imitieren.
- **Identitätsdiebstahl:** Betrüger nutzen Deepfake-Technologien, um die Identität einer echten Person zu fälschen. Damit können sie Schreibstil, Stimme, Sprechart, Erscheinungsbild oder sogar Körpersprache einer Person präzise imitieren. Solche Fälschungen können täuschend echt wirken und werden von Betrügern etwa zur Ausweisfälschung, in Video-Calls oder für „Schockanrufe" (siehe Infokasten) genutzt.
- **Fake-Profile:** Mit Deepfakes lassen sich auch „synthetische Identitäten" erschaffen. Diese Avatare werden oft zur Erstellung von Fake-Profilen genutzt, sogenannten Bots. Ziel ist es, diese Profile möglichst authentisch erscheinen zu lassen, um das Vertrauen echter Nutzer*innen zu gewinnen. Bots werden in sozialen Netzwerken eingesetzt, um mit echten Personen zu interagieren, Beiträge zu liken, zu kommentieren oder auch selbst zu erstellen. Mit Bots werden auch gefälschte Kundenprofile erstellt, um positive oder negative Bewertungen von Produkten oder Marken zu verbreiten. Das heißt über Bots wird versucht, Einfluss auf die öffentliche Wahrnehmung zu nehmen.
- **Verleumdung:** Durch Deepfakes können falsche und kompromittierende Informationen über Personen verbreitet werden, die ihrem Ruf (dauerhaft) schaden. Dies kann schwerwiegende Folgen auf das private, schulische oder berufliche Leben sowie die Psyche der Betroffenen haben. Besonders schwerwiegend ist hierbei, dass sich solche Inhalte online sehr schnell verbreiten und den Betroffenen damit meist keine Möglichkeit bleibt, diese zu widerlegen.

Schockanrufe

Bei Schockanrufen geben sich Kriminelle am Telefon als Familie oder Freunde der Zielperson aus, erklären in einer Notlage zu sein und dringend Geld zu brauchen. Durch den Einsatz von Stimm-Fakes, die der Zielperson vertraut klingen, kann sie eher geneigt sein, der Täuschung zu glauben und der Aufforderung nachzugehen.[13] Das Ziel von Schockanrufen ist, Panik und Angst bei der Zielperson auszulösen, damit diese überstürzt handelt und Geld überweist. Hier geht es zu Tipps im Umgang mit Schockanrufen.

→ **www.klicksafe.de/mnki00**

- **Missbrauch biometrischer Systeme:** Sich über biometrische Daten wie den Fingerabdruck, das Gesicht oder die Stimme zu authentifizieren, ist eine beliebte Methode, um sich Passwörter nicht merken zu müssen. Zudem gilt sie als sicher, um Zugriffsrechte zu Geräten, Netzwerken, Software oder auch physischen Räumen zu regeln. Doch immer realistischere Deepfakes stellen für biometrische Sicherheitskontrollen zunehmend ein Risiko dar. Vor allem wenn sie über das Telefon oder per Video erfolgen. Ein Weg, um dafür weniger anfällig zu sein, ist zum Beispiel die „Multifaktor-Authentifizierung". Das heißt neben biometrischen Daten werden auch noch andere Mechanismen genutzt, um den Zugang zu sensiblen Daten zu kontrollieren.

Desinformation und Demokratiegefährdung

Mit „Desinformation" sind nachweislich falsche Informationen gemeint, die gezielt erstellt und verbreitet werden, um andere zu täuschen oder in die Irre zu führen. Sprich: Es liegt eine **konkrete Täuschungsabsicht** vor. Auch dieses Phänomen ist nicht neu. Schon immer wurden Öffentlichkeiten getäuscht, um bestimmte Prozesse zu stören und in der Regel wirtschaftliche und politische Interessen durchzusetzen. **Neu** sind aber die **Geschwindigkeit, Quantität und Qualität von Desinformation**. Denn heute lassen sich durch generative KI massenhaft Texte, Bilder, Videos oder Audios erstellen, die täuschend echt wirken können. Alles, was es dazu braucht, sind meist nur wenige Klicks. Diese irreführenden Inhalte können über soziale Netzwerke (z. B. Bot-Netzwerke) schnell an Aufmerksamkeit gewinnen und große Reichweiten erlangen. Damit sind Deepfake-Technologien **mächtige Werkzeuge, die Lügen eine völlig neue Tragweite verleihen**.

Sie können zum Beispiel folgende Risiken mit sich bringen:

- **Einfluss auf Meinungsbildung:** Deepfakes können gezielt genutzt werden, um Lügen über bestimmte Personen, Gruppen oder Ereignisse zu verbreiten und so die öffentliche Meinung zu manipulieren. Zum Beispiel, indem Videos eine politische Persönlichkeit zeigen, die dort vermeintlich Dinge sagt oder tut, die sie in Wahrheit nie gesagt oder getan hat. Solche gefälschten Inhalten können in der Öffentlichkeit starke Gefühle wie Wut oder Hass hervorrufen oder Personen diskreditieren. Eine Taktik, die bestimmte Akteur*innen anwenden, um zum Beispiel vor Wahlen oder in Krisenzeiten gezielt Stimmung in der Gesellschaft zu machen oder Personen zu diskreditieren.
- **Extremismus:** Besonders gefährlich können Deepfakes sein, wenn extremistische Gruppen sie als Teil ihrer Online-Strategie nutzen. Denn darüber lassen sich irreführende bzw. falsche Narrative, extreme Positionen sowie Hass gegen bestimmte Personen(gruppen) verbreiten. Vor allem über visuelle Fakes können Emotionen transportiert und komplexe Zusammenhänge vereinfacht werden. Mithilfe von Deepfake-Tools können extreme Ideologien auch ein breiteres bzw. neues Publikum erreichen. So wurden zum Beispiel berüchtigte Reden von Adolf Hitler mithilfe von KI ins Englische übersetzt. Diese Fakes gingen viral und haben auf TikTok, X, Instagram und YouTube Millionen, vor allem junge Menschen erreicht.[14] Mit KI-Tools lassen sich Reden aber nicht nur übersetzen, sondern auch verändern. Solche falschen „Beweise“ können die Wahrhaftigkeit etablierter historischer Aufzeichnungen untergraben, indem sie Fakten aufweichen oder versuchen, die Geschichte neu zu erzählen. Etwa indem Hitler als „missverstandene“ Figur dargestellt bzw. verharmlost wird.
- **Manipulation von Wahlen:** Deepfakes werden auch bei Wahlkämpfen eingesetzt, um Desinformation über Politiker*innen in Umlauf zu bringen und Wähler*innen zu irritieren. Das kann dazu führen, dass politische Gegner*innen diskreditiert, geschwächt und vielleicht sogar aus den falschen Gründen nicht gewählt werden. Wie hoch jedoch der tatsächliche Einfluss von Deepfakes auf das Wahlverhalten der Bürger*innen in Demokratien ist, lässt sich nicht eindeutig sagen. Denn dieser Effekt kann nicht isoliert erfasst werden. Im Moment scheint es noch ein größeres Problem zu sein, dass Desinformation entsteht, wenn echte Medieninhalte aus dem Zusammenhang gerissen werden.[15]
- **Erosion von Wahrheit:** Wenn Menschen permanent mit Desinformation konfrontiert werden und rein theoretisch alles „Fake“ sein kann, was man online sieht und hört, kann das zu einer allgemeinen Skepsis gegenüber allen Informationen führen. Deepfakes können damit langfristig den Effekt haben, dass glaubwürdige Informationen entwertet und das Vertrauen in demokratische Säulen wie Medien, Wissenschaft, Wahlprozesse oder ein unabhängiges Rechtssystem zunehmend untergraben werden. Zudem können Deepfakes auch gezielt zum eigenen Vorteil genutzt werden. So können Personen des öffentlichen Lebens zum Beispiel behaupten, dass wahre Ereignisse und Informationen über sie selbst, die ihnen schaden können (z. B. ein Skandal), lediglich „Fake News“ oder „Deepfakes“ sind. Wird diese Lüge von der Öffentlichkeit geglaubt, kann die Person dadurch Vorteile erhalten, wie Glaubwürdigkeit, Einfluss und ein Ausbleiben der Konsequenzen. Dieses **Phänomen** wird in der Wissenschaft **„liar’s dividend“** genannt (dt.: Lügen-Dividende). Wenn Menschen

an einer verzerrten und falschen Wahrnehmung der Realität festhalten, selbst wenn sie um Deepfakes wissen, liegt das meist daran, dass die manipulierten Inhalte ihre eigenen Gefühle und Ansichten ansprechen. Dieses Verhalten lässt sich durch den sogenannten **Confirmation Bias** erklären. Demnach haben wir die Tendenz, Informationen zu bevorzugen, die zu unseren bestehenden Werten und Überzeugungen passen. Selbst dann, wenn sie falsch oder manipulativ eingesetzt worden sind. Je geschlossener und einfacher das Weltbild einer Person ist, umso eher entfaltet dieser Bestätigungsfehler seine Wirkung.

- **Gefährdung der Demokratie:** Demokratie benötigt den öffentlichen Diskurs und das Vertrauen der Bürger*innen in diesen. Das zunehmende Misstrauen kann jedoch dazu führen, dass sich Menschen immer mehr aus der demokratischen Öffentlichkeit zurückziehen. In „Echokammern" wenden sie sich dann vor allem jenen Informationen zu, die ihre Ansichten bestätigen. Deepfakes haben hier das Potential alternative, virtuelle Realitäten zu erschaffen, die von Interessensgruppen gezielt für ihre eigene Narrative eingesetzt werden. Solche Entwicklungen können Gesellschaften (weiter) spalten, die Polarisierung verstärken und autoritären Tendenzen mehr Raum und Einfluss verleihen.

Besonders **junge Menschen** befinden sich damit **in einem Dilemma**: Soziale Netzwerke sind für sie ein täglicher Begleiter und wichtige Informationsquelle, um sich über Nachrichten und das aktuelle Weltgeschehen zu informieren. Sie sind jedoch auch wichtiger Dreh- und Angelpunkt, über den sich Desinformation, auch in Form von Deepfakes, verbreitet. Gleichzeitig fehlt jungen Menschen oft die Erfahrung, Informationen im Netz auf ihre Glaubwürdigkeit zu prüfen. Eine Sonderauswertung der PISA-Studie 2022 zeigt, dass viele Jugendliche in Deutschland Probleme damit haben, „Fake News" zu erkennen und die Qualität von Online-Informationen zu beurteilen. Viele prüfen auch nicht, ob Informationen aus dem Internet korrekt sind, bevor sie diese in sozialen Medien mit anderen teilen. Wenn junge Menschen nicht wissen, wie sie echte und gefälschte Informationen unterscheiden bzw. Informationen auf Glaubwürdigkeit hin prüfen können, kann das nachhaltige Folgen haben. Denn jüngere Menschen, die noch auf der Suche nach der eigenen Identität und in ihrem (politischen) Weltbild nicht gefestigt sind, können durch Manipulationsstrategien und Desinformation nachhaltig beeinflusst werden.

„Voting this tuesday only enables the republicans in their quest to elect donald trump again. Your vote makes a difference in november, not this tuesday."

Robocalls mit Deepfake-Biden

Im Zuge des US-Wahlkampfs 2024 lösten automatisierte Anrufe, sogenannte Robocalls, große Sorge aus, die sich an demokratisch gesinnte Bürger*innen im US-Staat New Hampshire richteten. Dort gab sich eine KI-generierte Stimme als Joe Biden aus, damals noch amtierender US-Präsident, die ihm zum Verwechseln ähnlich war. Der vermeintliche Biden forderte dort die Wähler*innen auf, sich an den Vorwahlen im Bundesstaat nicht zu beteiligen, um angeblich die Wiederwahl Trumps durch die Republikaner*innen zu verhindern. Ziel des automatisierten Anrufs mit KI-generierter Stimme war, die Menschen gezielt in die Irre zu führen. Das Bild wurde mithilfe von KI generiert.
Quelle: GPT 4o, OpenAI (17.4.2025)

MEHR INFOS

Es ist wichtig, dass junge Menschen bestimmte Kompetenzen erlernen, um wahre von falschen Informationen unterscheiden zu können. Unter diesen beiden Links finden Sie weitere Informationen, Tipps und Handlungsmöglichkeiten für Jugendliche, pädagogische Fachkräfte und Eltern zu den Themen:

→ **www.klicksafe.de/mnme00**
→ **www.klicksafe.de/mne00**

Deepfakes und extremistische Inhalte

Auf TikTok verbreitete sich dieses Sharepic im Stil eines authentisch wirkenden Disney-Pixar-Filmplakates. Das Bild wurde mit KI erstellt. Der Titel des angeblich beworbenen Films „Caust" ist eine Abkürzung für Holocaust. Das Deepfake-Bild zeigt ein klares Spannungsverhältnis zwischen Grausamkeit und Verniedlichung durch den Pixar-typischen Comic-Stil und soll provozieren. „Humor" wird hier gezielt als Mittel eingesetzt, um extreme Botschaften und Figuren herunterzuspielen und zu normalisieren. *Quelle: TikTok (2024). Weitere Informationen: www.jugendschutz.net/mediathek/artikel/kurz-analyse-disney-pixar-challenge*

MANIPULATION DURCH BILDER – EIN ALTES PHÄNOMEN UND MÄCHTIGES WERKZEUG

Bekannt sind Redewendungen wie „Ein Bild sagt mehr als tausend Worte", „Bilder bleiben im Kopf" oder „Ich glaube es erst, wenn ich es sehe!". Manche behaupten auch, der Mensch sei ein „Augentier". In allen Aussagen steckt ein wahrer Kern. Denn der Sehsinn zählt zu einem unserer zentralsten Sinnesorgane, über das wir unsere Umgebung wahrnehmen. Das Gehirn reagiert auf **Bilder** sogar um ein Vielfaches schneller als auf Texte, was sie zu **wichtigen Informationsträgern** macht. Über Bilder lassen sich Geschichten erzählen und subtile Botschaften transportieren. Zudem können Bilder in uns Gefühle auslösen bzw. verstärken. Besonders wenn etwas Negatives dargestellt wird, wie Gewalt oder Angst. Das liegt daran, wie Menschen Informationen verarbeiten. Denn wir schenken negativen Reizen oft mehr Beachtung als positiven und können uns auch besser an sie erinnern.

In der Wissenschaft ist das Phänomen auch als **„Negativitätsbias"** bekannt. Demnach wird in uns ein uraltes Reaktionsmuster aktiviert, das bei unseren Vorfahren für lange Zeit das Überleben sicherte. Denn es überlebten vor allem die Menschen, die die Anzeichen von Gefahr rechtzeitig erkannten und schnell reagierten. Diese negative Verzerrung ist im Journalismus auch unter dem Slogan „Bad news are good news!" bekannt. Denn mit dem Ringen um die mediale Aufmerksamkeit haben **Emotionen** an Bedeutung gewonnen. Demnach lassen sich mit negativen Schlagzeilen, sprich Krisen, Kriegen und Katastrophen Menschen besser erreichen. Doch Menschen sind Medieninhalten nicht einfach ausgeliefert und nehmen diese passiv auf. **Rezipienten und Massenmedien** stehen vielmehr in einer **Wechselwirkung** zueinander (vgl. Uses-and-Gratifications-Ansatz). Jeder verfügt über seine persönlichen Normen, Werte und Bedürfnisse, die wie eine Art Filter wirken. Soll heißen: Wir beeinflussen ebenfalls, welchen Inhalten und Medien wir uns aktiv zuwenden und welche Bedeutung wir ihnen dann zuschreiben. Ob eine intendierte Wirkung durch einen Medieninhalt tatsächlich eintritt, ist also kein Naturgesetz.

Auch der Blick in die Vergangenheit zeigt: Die Manipulation von Medieninhalten ist **kein neues Phänomen** und das **Wissen um die Macht der Bilder** reicht weit in unsere Menschheitsgeschichte zurück. Schon in der Antike nutzte man zum Beispiel Bildprägungen auf Geldmünzen als ein mächtiges Werkzeug für politische Propaganda, um darüber Macht, Werte und wichtige Botschaften an die Bevölkerung zu vermitteln. Und auch lange vor dem Zeitalter der KI-Technologien wurden Bilder manipuliert, um wirtschaftlich und politisch Einfluss zu nehmen.

Wie mit Deepfakes umgehen?

Hör auf dein Bauchgefühl!

Sinne schärfen und Medieninhalte bewusst konsumieren: Soziale Medien sind sehr dynamisch. Inhalte werden dort rund um die Uhr abgerufen und in Sekundenschnelle geteilt. Die perfekte Umgebung für Deepfakes, die sich so in Windeseile verbreiten können. Hört oder sieht man nicht genau hin, können sie auch schnell mal als „echt“ durchgehen. Umso wichtiger ist es, sich Zeit zu nehmen, wenn man Medieninhalten begegnet. **Achte auf die Quelle, den Inhalt, Kontext sowie sichtbare oder hörbare Details. Und: Höre vor allem auf dein Bauchgefühl!** Erfolgt dies regelmäßig, schärfst du damit die eigene Sinne und kannst Fehler und Ungewöhnliches besser erkennen.

Mach den Emotions-Check!

Deepfakes triggern „deep feelings“: Starke Gefühle erregen Aufmerksamkeit und sorgen dafür, dass wir Inhalten eher Beachtung schenken und sie mit anderen teilen. Vor allem wenn sie negativ sind. Doch Vorsicht! Emotionen sind auch der Treibstoff für Desinformation. Vor allem das Visuelle bietet hierfür gute Anknüpfungspunkte, da ein Mangel an Inhalt, Kontext und Argumenten herrscht. Bilder lassen sich leicht emotional aufladen und mit subtilen Codes versehen. Deepfakes werden daher oft als Mittel genutzt, um solche irreführenden Medieninhalte zu verbreiten.

MATERIAL

Das Plakat „Achtung Deepfakes!“ enthält diese Tipps nochmal in Kürze, damit Schüler*innen wissen, wie man Deepfakes entlarven und sich vor gezielter Manipulation schützen kann. Das Plakat richtet sich an Kinder und Jugendliche zwischen 10 und 14 Jahren.

→ **www.klicksafe.de/materialien/achtung-deepfakes**

Achte auf Fehler!

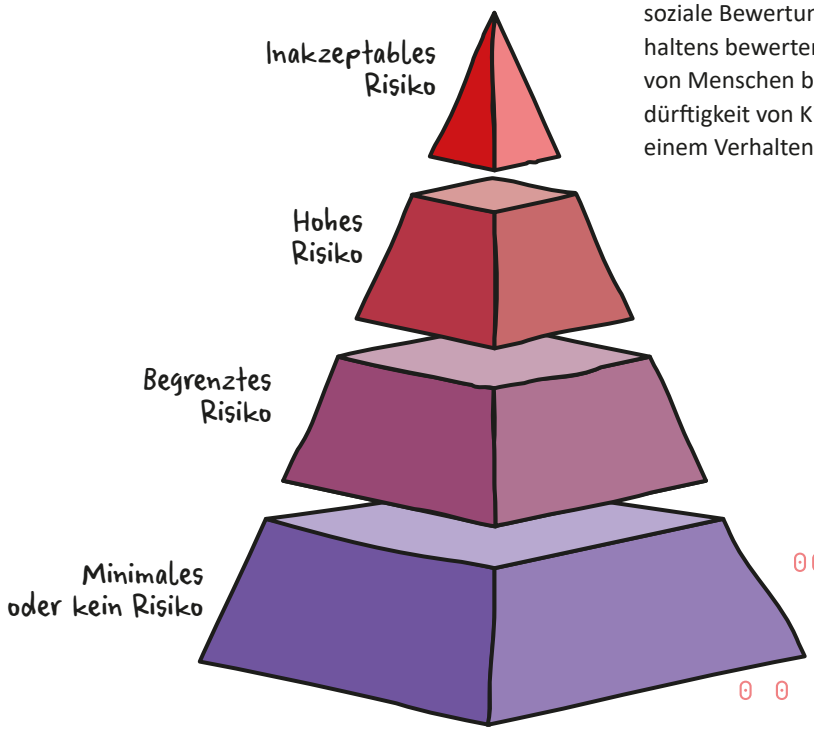
Der „Teufel“ steckt im Detail: Deepfakes sind nicht frei von technischen Fehlern. Doch man muss zum Teil schon gut Hinsehen und Hinhören, um sie als solche zu erkennen. Denn die Technologien dahinter werden immer besser und der „Teufel“ steckt damit zunehmend im Detail. Suche daher gezielt nach visuellen und hörbaren Hinweisen. Bei der **Bildqualität** kann zum Beispiel eine unnatürliche oder zu perfekte Beleuchtung auffällig sein, sowie Verzerrungen, verschwommene Stellen oder sinnlose Schatten und Flecken. Auch **Mimik** und **Körper** von Personen können Hinweise geben. Etwa durch unnatürliche Bewegungen (z. B. bei Körper, Augen, Lippen), unpassende Proportionen des Körpers, Unschärfen rund um das Gesicht oder im Seitenprofil, fehlende Details (z. B. bei Zähnen, Augen), seltsame Formen (z. B. bei Gliedmaßen, Ohren, Nase) oder durch zu perfekt wirkende Gesichter. Auch auf **akustische Hinweise** ist zu achten, wie zum Beispiel Hintergrundgeräusche, unnatürlicher Klang der Stimme, seltsame Betonungen und Sprachmuster sowie ein nicht synchroner Ton zur Mundbewegung.

Bleib skeptisch!

Vorsicht bei unbekannten und nicht journalistischen Quellen: Generell sollte man Informationen, vor allem politische, **immer kritisch hinterfragen und sich über die Glaubwürdigkeit des Inhalts vergewissern**. Prüfe hierfür, wer der Absender ist, welche Absicht hinter der Nachricht steckt und ob die dargelegten „Fakten“ stimmen können. Stammt die Information aus einer dir unbekannten bzw. nicht journalistischen Quelle, solltest du nach mehr Informationen zum Thema suchen und prüfen, ob seriöse Medien darüber berichten. Du kannst auch Faktenchecker nutzen, um herauszufinden, ob das Ereignis (so) tatsächlich stattgefunden hat. Auch eine Bilder-Rückwärtssuche kann dir dabei helfen, einen Deepfake zu entlarven.

Wie ist die Rechtslage?

Der **AI Act der EU**, auch KI-Gesetz genannt, ist der weltweit erste umfassende **Rechtsrahmen, der den Einsatz von künstlicher Intelligenz (KI) reguliert**. Die Verordnung ist am 1. August 2024 in Kraft getreten und erlangt in vier Stufen bis August 2027 Geltung. Das Gesetz richtet sich an Anbieter (z. B. Entwickler) und Bereitsteller von KI-Systemen und ist Teil des digitalpolitischen Maßnahmenpakets der EU. Mit der Verordnung will die EU auch den **Herausforderungen durch Deepfakes begegnen**.
Ziel des Gesetzes soll sein, vertrauenswürdige KI-Systeme in Europa zu fördern und Chancen zu nutzen, die durch KI-Technologien entstehen können. Wie gesellschaftliche Vorteile, Förderung von Innovation und globale Wettbewerbsfähigkeit. Gleichzeitig sollen Risiken minimiert werden, die durch KI-Systeme für Bürger*innen sowie die Gesamtgesellschaft entstehen können. Der Einsatz von KI soll sicher, rechtmäßig, diskriminierungsfrei und ethisch vertretbar erfolgen.
Der AI Act verfolgt dabei einen **risikobasierten Ansatz**. Das bedeutet, dass KI-Systeme je nach Risiko unterschiedlichen Verpflichtungen unterliegen. Die KI-Verordnung unterscheidet dabei **vier Risikostufen**:[16]



1 Minimales/kein Risiko:
Hier eingestufte KI-Systeme unterliegen keinen speziellen Regeln. Die meisten KI-Anwendungen fallen in diese Kategorie (z. B. Video-Spiele oder Spam-Filter).

2 Begrenztes Risiko:
Hier eingestufte KI-Systeme unterliegen einer Transparenzpflicht. Anbieter müssen sicherstellen, dass Nutzende wissen, dass sie mit einem KI-System interagieren (z. B. Chatbot) oder mit einem KI-generierten Inhalt konfrontiert werden (z. B. Deepfake).

3 Hohes Risiko:
Hier eingestufte KI-Systeme werden streng kontrolliert, da ihre Anwendung schwerwiegende Risiken für Gesundheit, Sicherheit und Grundrechte darstellen können. Dazu zählen zum Beispiel KI-Systeme in den Bereichen Verkehr, Bildung, Medizin oder Strafverfolgung. Bei solchen Hochrisiko-KI-Systemen muss eine wirksame menschliche Aufsicht sichergestellt sein.

4 Inakzeptables Risiko:
Hier eingestufte KI-Systeme sind verboten, da sie als eindeutige Bedrohung für die Sicherheit, den Lebensunterhalt und die Rechte der Menschen gelten. Dazu gehören zum Beispiel soziale Bewertungssysteme, die Personen aufgrund ihres Verhaltens bewerten. Oder manipulative KI, die das Verhalten von Menschen beeinflusst. Und auch KI, die die Schutzbedürftigkeit von Kindern und Jugendlichen ausnutzt, um sie zu einem Verhalten zu bewegen, dass ihnen Schaden zufügt.

Weiterlesen

Mehr zum Thema bei klicksafe

- **klicksafe Quiz „Deepfakes und Co.“:** Mit dem Quiz können Jugendliche testen, wie gut sie über Deepfakes und Co. Bescheid wissen und auf spielerische Weise echtes Faktenwissen erlernen. Geeignet für 5. bis 7. Klasse. Das Quiz ist in Kooperation mit ZDF logo! entstanden.

→ **www.klicksafe.de/materialien/quiz-zum-thema-deepfakes**

- **klicksafe Quiz „Deepfake Detectives“:** Mit dem Quiz können Jugendliche testen, ob sie einen Deepfake erkennen oder nicht. Geeignet ab 7.Klasse.

→ **www.klicksafe.de/materialien/quiz-deepfake-detectives**

- **klicksafe Plakat „Achtung Deepfakes“:** Auf dem Plakat finden sich schnelle und einfache Tipps, wie man Deepfakes entlarven und sich vor Manipulation durch gefälschte Fotos und Videos schützen kann. Geeignet für 5. bis 7. Klasse.

→ **www.klicksafe.de/materialien/achtung-deepfakes**

- **Vierteiliger Expert*innen-Talk mit Dr. Bernd Zywietz zum Thema „Deepfakes und Extremismus“:** Deepfakes können besonders gefährlich sein, wenn sie zum Zwecke gezielter Desinformation von extremistischen Gruppen erstellt und verbreitet werden. Zum Beispiel, um politisch Stimmung zu machen.

→ **www.klicksafe.de/news/expertinnen-talk-zum-thema-deepfakes-und-extremismus**

- **klicksafe Material „Ethik macht klick. Meinungsbildung in der digitalen Welt“:** Das Material ist in Zusammenarbeit mit dem Institut für digitale Ethik an der HdM Stuttgart entstanden. Es bietet Pädagog*innen Einblicke in das Informationsverhalten von Jugendlichen, gibt Hilfestellung beim Analysieren und Erkennen von Desinformationsstrategien und zeigt Auswirkungen von Falschinformation für die demokratische Gesellschaft auf.

→ **www.klicksafe.de/materialien/ethik-macht-klick-meinungsbildung-in-der-digitalen-welt**

- **klicksafe Material „Rechts. Extrem. Online. Wie man Jugendliche gegen rechtsextreme Einflüsse im Internet stark macht“:** Rechtsextreme Akteur*innen nutzen das Internet gezielt, um junge Nutzer*innen zu erreichen und für sich zu gewinnen. Das Unterrichtsmaterial befähigt junge Menschen, rechtsextreme Propaganda zu erkennen und sich mit den Herausforderungen rechtsextremer Narrative im Netz kritisch und selbstbestimmt auseinanderzusetzen. Passend zum Material gibt es auch das Actionbound-Spiel „#cleanyournetwork“. Dort lernen Jugendliche, wie sie rechtsextreme Online-Strategien entlarven und sich dagegenstellen können.

→ **www.klicksafe.de/materialien/rechts-extrem-online-wie-man-jugendliche-gegen-rechtsextreme-einfluesse-im-internet-stark-macht**

→ **www.klicksafe.de/materialien/actionbound-cleanyournetwork-bootcamp-gegen-rechtsextremen-hass-und-fuer-demokratie-auf-social-media**

- **klicksafe Arbeitsblatt „Willst du mit mir Fakten checken gehen?“:** klicksafe und die Medienscouts NRW bieten Pädagog*innen für den Einsatz in der Peer-to-Peer-Arbeit ein Begleitmaterial an. Jugendliche können unter dem Titel „Willst du mit mir Fakten checken gehen?“ damit Gleichaltrigen den richtigen Umgang mit Desinformationen beibringen. Das Material enthält auch einen Link zum Actionbound-Spiel „Im Bunker der Lügen“. Dort können Jugendliche ihr Wissen zu Fake News und Verschwörungserzählungen testen bzw. vertiefen.

→ **www.klicksafe.de/materialien/begleitmaterial-willst-du-mit-mir-fakten-checken-gehen**

- **klicksafe Themenbereich „Deepfakes“**

→ **www.klicksafe.de/desinformation-und-meinung/deepfakes**

- **klicksafe Themenbereich „Künstliche Intelligenz“**

→ **www.klicksafe.de/kuenstliche-intelligenz**

- **klicksafe Themenbereich „Sexualisierte Gewalt durch Bilder“**

→ **www.klicksafe.de/sexualisierte-gewalt-durch-bilder**

Weitere Links zum Thema

- **jugendschutz.net Jahresbericht:** Der Jahresbericht erfasst das Gefährdungspotenzial für Kinder und Jugendliche im Netz und nimmt dabei (ab 2023) auch den wachsende Einfluss generativer KI in den Fokus, der die Verbreitung problematischer Inhalte zusätzlich befeuert.

→ **www.jugendschutz.net/mediathek**

- **JUUUPORT** hat das **Video „Täuschend echt! Was du über Deepfakes wissen solltest."** veröffentlicht, das junge Menschen für die damit verbundenen Risiken sensibilisiert und ihnen konkrete Tipps gibt.

→ **www.juuuport.de/infos/news/taeuschend-echt-video-ueber-deepfakes**

- **FINNreporter KieKI:** Im Projekt „KieKI – Kinder erklären KI" entdecken die Kinderreporter von fragFINN die Welt der künstlichen Intelligenz mit Kinderaugen und -ohren. Es gibt dort zahlreiche Interviews mit KI-Expert*innen, selbstgedrehte Erklär-Filme und Quizze – auch zum Thema „Deepfakes". Das Angebot richtet sich an Kinder ab 8 Jahren.

→ **https://reporter.fragfinn.de**

- Das informative **Comic-Essay „Schokoroboter & Deepfakes"** thematisiert KI aus der Perspektive von Jugendlichen.

→ **https://schokofakes.ai**

- Das **Landesmedienzentrum Baden-Württemberg** stellt in einer Datenbank insgesamt **21 Lernideen zum Thema KI** kostenlos zur Verfügung, die aufzeigen, wie man die Potenziale von KI in der Bildung einsetzen kann.

→ **www.lmz-bw.de/21-ki-lernideen**

- Der **Bayerische Rundfunk** stellt online eine **Unterrichtseinheit zu Deepfakes** zur Verfügung, adressiert an die Sekundarstufe. Unter dem Link gibt es auch einen niedrigschwelligen Wissenstest für Schüler*innen.

→ **www.br.de/sogehtmedien/stimmt-das/deepfakes/index.html**

- Über die Plattform **SchulKI** kann man einen **Chatbot für den Unterricht** nutzen, der mit seinen Features und in puncto Datenschutz speziell für den Bildungsbereich entwickelt wurde. Über den Anbieter können auch Bilder generiert werden.

→ **https://schulki.de**

- Auf der Plattform **digital.learning.lab** teilen Lehrkräfte ihre **Erfahrungen und ihr Wissen über den Einsatz von KI im Unterricht**. Das Angebot soll Lehrkräfte dabei helfen, selbst digitale Kompetenzen zu erwerben oder aufzubauen, um Schüler*innen in ihrer Kompetenzentwicklung gut zu unterstützen. Neben Trends und einer umfangreichen Toolbox gibt es dort „Good Practice"-Beispiele in Form von Unterrichtsbausteinen.

→ **https://digitallearninglab.de**

- Das ausgezeichnete **Projekt deepfake detective** zielt mit seinen interaktiven Workshops darauf ab, Schüler*innen zum Thema Deepfakes zu sensibilisieren und in ihrer Medienkompetenz zu fördern. Ein Highlight des Workshops: Bei einer Station kommt Virtual Reality zum Einsatz. Schüler*innen absolvieren eine „Deepfake Detective"-Ausbildung, um ihre Wahrnehmung und den Blick für Auffälligkeiten in unterschiedlichem Videomaterial zu trainieren. Der Workshop ist öffentlich gefördert und daher kostenfrei. Die einzigen Kosten sind die An- und Abreisekosten.

→ **https://deepfake-detective.de**

Andere Deepfake Quizze

→ **www.saferinternet.at/news-detail/fight-fakes-neue-quiz-zum-thema-deepfakes**

→ **https://fakes.thraets.org**

→ **https://deepfact.3duniversum.com/quiz**

→ **https://quiz.iproov.com**

Übersicht über die Projekte

	Projekt 1	Projekt 2	Projekt 3
Titel	Deepfakes auf der Spur – Fakes verstehen, prüfen und erkennen	Die Macht der Bilder – Risiken durch Deepfakes verstehen	Jetzt wird's deep – Deepfakes selbst erstellen
Kurzbeschreibung	Die SuS lernen verschiedene Formen von Deepfakes kennen. Anhand einer Checkliste entscheiden sie anschließend, ob es sich bei den gezeigten Medieninhalten um einen Deepfake handelt oder nicht.	Die SuS befassen sich mit dem Phänomen „Bild- und Videomanipulation". Sie reflektieren die negativen Konsequenzen, die hierdurch für einzelne Personen, Menschengruppen oder die Gesamtgesellschaft entstehen können.	Die SuS sollen einen eigenen Deepfake erstellen. Anhand vorgegebener Kriterien sollen verschiedene KI-Tools miteinander verglichen werden.
Lernziele	Die SuS können erklären, was ein Deepfake ist. Die SuS lernen anhand welcher Merkmale sie Deepfakes erkennen können. Die SuS lernen verschiedene Formen von Deepfakes kennen.	Die SuS erkennen, dass KI-manipulierte bzw. KI-generierte Medieninhalte (Text, Bild, Video und Audio) missbräuchlich genutzt werden können. Sie lernen verschiedene Missbrauchsszenarien kennen und können diese benennen.	Die SuS können ein eigenes Bild mithilfe eines KI-Generators erstellen. Die SuS lernen die Grundlagen des Promptens kennen und können gezielt Prompts formulieren. Die SuS lernen, Deepfake-Tools kreativ und verantwortungsvoll zu nutzen.
Unterrichtsstunden á 45 Minuten	1–2	1–2	2–3
Methoden und Material	Tafelsturm und theoretische Einführung „Was sind Deepfakes?" Bildvergleich und Analyseübung mit dem Pdf „Spot the Fake" Spielerisches Lernen mit Quiz „Deepfake Detectives" und Checkliste „Deepfake Detectives"	Multimediale Impulse anhand eines YouTube-Videos („Historische Bildmanipulation – Photoshop der Geschichte") bzw. einer Bildreihe („Klassiker der Bildmanipulation") Kooperative Gruppenarbeit mit dem AB „Deep Fake – Deep Impact" Recherchieren und Entdecken mithilfe einer Linksammlung	Aktivierender Einstieg über Bilder-Rätsel Erkundung und Anwendung von KI-Tools zur Bildgenerierung (z. B. über OpenAI (DALL-E oder GPT 4o), Fobizz, SchulKI oder paddy)* Experimentieren mit Prompts und Vertiefung durch KI-gestütztes „Superprompting" Kreative Gruppenarbeit mit dem AB „Jetzt wird's deep" zur Erstellung einer Mini-Kampagne
Zugang zu Internet/PC	ja	ja	ja
Klassenstufe	Empfohlen ab 7./8. Klasse	Empfohlen ab 7./8. Klasse	Empfohlen ab 7./8. Klasse

Projekt 1

Deepfakes auf der Spur – Fakes verstehen, prüfen und erkennen

Kurzbeschreibung	Die SuS lernen verschiedene Formen von Deepfakes kennen. Anhand einer Checkliste entscheiden sie anschließend, ob es sich bei den gezeigten Medieninhalten um einen Deepfake handelt oder nicht.
Lernziele	▪ Die SuS können erklären, was ein Deepfake ist. ▪ Die SuS lernen anhand welcher Merkmale sie Deepfakes erkennen können. ▪ Die SuS lernen verschiedene Formen von Deepfakes kennen.
Unterrichtsstunden á 45 Minuten	1–2
Methoden und Material	▪ Tafelsturm und theoretische Einführung „Was sind Deepfakes?“ ▪ Bildvergleich und Analyseübung mit dem Pdf „Spot the Fake“ ▪ Spielerisches Lernen mit Quiz „Deepfake Detectives“ und Checkliste „Deepfake Detectives“
Zugang zu Internet/PC	ja
Klassenstufe	Empfohlen ab 7./8. Klasse

Was sind Deepfakes?

Bei Deepfakes handelt es sich um **realistisch wirkende Medieninhalte** (z. B. Bilder, Videos oder Audios), die **durch künstliche Intelligenz verändert oder sogar komplett neu erzeugt** worden sind. Mit **„Deep“** ist gemeint, dass hier „Deep Learning“ zum Einsatz kommt. Eine spezielle Lernmethode, bei der Computer lernen, Muster und Zusammenhänge in sehr großen Datenmengen zu erkennen. Sie lernen zum Beispiel, wie das Gesicht einer Person aussieht, wie sich ihre Mimik verhält, wie ihre Stimme klingt oder in welchem Schreibstil sie E-Mails verfasst. Mit **„Fake“** ist gemeint, dass es sich bei den Inhalten um eine Fälschung bzw. einen computergenerierten Inhalt handelt.

Einstieg

Öffnen Sie über den Link das **Pdf „Spot the Fake!“**, das sowohl Deepfakes als auch „echte“ Bilder enthält. Zeigen Sie der Klasse nacheinander die Inhalte und fragen Sie die SuS nach ihrer Einschätzung zu jedem Bild: Fake or not?
→ **www.klicksafe.de/spot-the-fake**

Lassen Sie die SuS ihre Entscheidung direkt bzw. intuitiv treffen, ohne dass das Bild lange analysiert wird. Der erste Eindruck zählt – ähnlich wie beim Scrollen von Inhalten in Social Media. Dementsprechend soll die Bewertung der SuS erstmal keine Erklärung enthalten, sondern nur deutlich machen, ob sie den gezeigten Inhalt für echt halten oder nicht. Eine Auflösung der Ergebnisse erfolgt später.

Die Bewertung kann zum Beispiel so abgefragt werden:

Fake	**Echt**
klatschen	nicht klatschen
„Fake“ rufen	stumm bleiben
Arm hoch	Arm bleibt unten

Fragen an die SuS:

- Woran merkt ihr, dass es sich vermutlich um einen „Deepfake“ handelt?
- Gab es Anzeichen oder Merkmale, die ihr entdecken konntet?
- Gab es Inhalte, bei denen euch die Entscheidung eher schwer gefallen ist? Wenn ja, warum?
- Seid ihr schon mal manipulierten/gefälschten Medieninhalten (z. B. Audios, Bilder oder Videos) im Internet begegnet, die vermutlich mit künstlicher Intelligenz erstellt wurden?
- Falls ja: Wo und in welchem Zusammenhang?

Sammeln Sie mit den SuS nun mündlich oder schriftlich an der Tafel, welche **Merkmale** für sie auf einen **„Fake“** hingedeutet haben. Besprechen Sie die genannten Argumente und rufen damit das Vorwissen der SuS ab. Diskutieren Sie anschließend, ob es Inhalte gab, bei denen die SuS eindeutige Anzeichen für einen möglichen „Fake“ erkennen konnten. Oder ob es ihnen bei manchen Inhalten schwerfiel, dies zu bewerten. Ergänzen Sie ggf. neu genannte Merkmale. Fragen Sie die SuS auch zu ihren Erfahrungen mit Deepfakes.

Auflösung:

Präsentieren Sie die Auflösung und besprechen Sie Ergebnisse, bei denen es abweichende Einschätzungen gab bzw. mit denen die SuS nicht gerechnet haben.

Einstiegs-Alternative:

Sie können die SuS befragen, was ihnen spontan einfällt, wenn sie den Begriff „Deepfakes“ hören. Erstellen Sie hierfür einen **Tafelsturm** zu dem Begriff „Deepfake“, um die SuS zu aktivieren. Schreiben Sie den Begriff vor der Stunde an (Tafel, Flipchart, Präsentations-PC, etc.). Die SuS werden aufgefordert, ihr Vorwissen stichpunktartig zu nennen. Schreiben Sie die Beiträge auf die Präsentationsfläche. In einem nächsten Schritt können Sie die SuS fragen, ob sie wissen, was mit dem Wortteil „Deep“ im Begriff Deepfake gemeint sein könnte. Führen Sie die genannten Stichpunkte in einer Definition zusammen, die Sie den SuS mündlich vermitteln oder visuell präsentieren.

Einstiegs-Alternative für höhere Klassenstufen:

Vertiefen Sie **vorhandenes Wissen der SuS** und zeigen Sie die verschiedenen Formen von Deepfakes. Falls ausreichend Zeit vorhanden ist, erklären Sie ggf. die Technologien dahinter. Mehr Informationen finden Sie in den Kapiteln: „Welche Formen von Deepfakes gibt es?“ (s. Seite 9) und „Wie werden Deepfakes erstellt?“ (s. Seite 6).

Erarbeitung

Gruppenarbeit „Deepfake Detectives“:

Teilen Sie die **Checkliste „Deepfake Detectives“** aus, die wichtige Merkmale auflistet, woran sich Fakes erkennen lassen. Alternativ können Sie die Liste auch frontal auf der Präsentationsfläche anzeigen, so dass sie jeder sehen kann.

Die SuS sollen sich in Kleingruppen zusammenfinden und das **Quiz „Deepfake Detectives“** in Ruhe durchgehen. Die Übung kann auch frontal mit der ganzen Klasse durchgeführt werden. Die SuS sollen im Gespräch mit der Gruppe festlegen, anhand welcher Merkmale das jeweils angezeigte Beispiel „Fehler“ bzw. „Merkmale” für einen Deepfake enthält. Die Checkliste dient als Unterstützung bei der Entscheidungsfindung.

→ **www.klicksafe.de/materialien/quiz-deepfake-detectives**
→ **www.klicksafe.de/checkliste-deepfake-detectives**

Sicherung

Auswertung der Aufgaben:

Besprechen Sie die Quiz-Ergebnisse sowie die Stellen, an denen die SuS Schwierigkeiten hatten, die Deepfakes als solche zu erkennen. Fassen Sie anschließend nochmal die wichtigsten Punkte entlang der folgenden Fragen zusammen.

Fragen an die SuS:

- Warum sehen Deepfakes oft so „real“ aus?
- Welche Formen von Deepfakes gibt es?
- Woran könnt ihr Deepfakes erkennen?
- Was könnt ihr tun, wenn ihr einem Deepfake online begegnet?

Sammeln Sie mit den SuS mündlich oder schriftlich Ideen, wie sie am besten mit „Fakes“ online umgehen sollten (s. Infokasten).

Umgang mit Deepfakes

- Inhalte nicht einfach weiterleiten!
- Beitrag melden! Zum Beispiel in sozialen Netzwerken oder bei Meldestellen.
- Andere warnen! Zum Beispiel, indem man über die Kommentarfunktion Inhalte als „Fake“ kennzeichnet.
- Faktenchecker nutzen! Zum Beispiel Angebote von Mimikama, Correctiv, DPA- oder AFP-Faktencheck.
- Zusammenhang prüfen! Zum Beispiel: Wer ist zu sehen/zu hören und was wird behauptet? Passt das, was man sieht und hört, zusammen? Ist es wahrscheinlich, dass diese Person so etwas sagt bzw. sich so verhält? Etc.

Weitere Informationen dazu, wie man mit Deepfakes umgehen kann, finden Sie hier:
→ **www.klicksafe.de/mmdiei5**

Idee: Hängen Sie das **Plakat „Achtung Deepfakes“** im Klassenzimmer auf und besprechen ggf. nicht genannte Punkte. Die SuS können auch noch ein passendes Quiz spielen, das in Kooperation mit ZDF logo! entstanden ist.

→ **Quiz: www.klicksafe.de/materialien/quiz-zum-thema-deepfakes**
→ **Plakat Download und Bestellung: www.klicksafe.de/materialien/achtung-deepfakes**

DEEPFAKE DETECTIVES
CHECKLISTE

Bildqualität:

- Schatten und Licht wirken (vor allem bei Bewegung) unnatürlich bzw. zu perfekt
- Bildfehler sind vorhanden wie verschwommene Stellen, deplatzierte Flecken, falsche Buchstaben bzw. fehlerhafte Texte, verzerrte Objekte oder Personen im Hintergrund, etc.
- Qualität des Bildes verändert sich beim Hineinzoomen

Kontext:

- Objekte oder Personen im Hintergrund passen inhaltlich nicht zum Gesamt-Setting
- Details wie Muttermale, Tattoos oder andere sichtbare Besonderheiten einer Person sind im Vergleich zu älteren Originalaufnahmen nicht vorhanden/hinzugekommen/an der falschen Stelle
- Aussagen und Handlungen passen nicht zum üblichen Verhalten der Person
- Betonung, Sprachmuster und Körperhaltung passen nicht zum üblichen Verhalten der Person
- Zeit und Ort der Aufnahme sind nicht nachvollziehbar
- Ort der Aufnahme ist über Google Maps nicht auffindbar/sieht in Wirklichkeit anders aus

Mimik und Körper:

- Gesicht, Haut oder Haare wirken unnatürlich (z. B. zu glatt, zu perfekt)
- Augen, Mund oder Zähne sind nicht detailliert/trennscharf
- Übergänge rund um das Gesicht sind unstimmig/verschwommen
- Ohren, Nase oder Zähne haben seltsame Formen
- Gliedmaßen sind unnatürlich (z. B. Form, Anzahl Finger, etc.)
- Pupillen nicht kreisrund
- Körperproportionen wirken unpassend
- Alter von Gesicht und Körper passen nicht zusammen

Bewegung (bei Video-Deepfakes):

- Augenbewegung ungewöhnlich, keine Veränderung der Pupillen (Durchmesser ändert sich nicht)
- Lippenbewegungen wirken unnatürlich
- Gesicht wird in der Seitenansicht unscharf
- begrenzte oder unnatürliche Bewegung von Gesicht, Kiefer, Falten, Grübchen
- Körperbewegung ist eingeschränkt bzw. wirkt unnatürlich
- Bewegung von Händen/Armen ist beim Sprechen eingeschränkt bzw. wirkt unnatürlich
- Übergänge im Bildmaterial unscharf oder „holprig“
- Bildqualitäten innerhalb eines Videos unterschiedlich

Ton (bei Audio- und Video-Deepfakes):

- Audioqualität ist schlecht bzw. unnatürlich
- Stimme und Mundbewegungen sind nicht synchron
- Betonung der Stimme ist monoton bzw. wirkt unnatürlich
- Aussprache von Wörten wirkt „komisch“ oder klingt falsch
- beim Sprechen gibt es unnatürliche Pausen/Verzögerungen
- metallischer Sound

Projekt 2

Die Macht der Bilder - Risiken durch Deepfakes verstehen

Kurzbeschreibung
Die SuS befassen sich mit dem Phänomen „Bild- und Videomanipulation“. Sie reflektieren die negativen Konsequenzen, die hierdurch für einzelne Personen, Menschengruppen oder die Gesamtgesellschaft entstehen können.

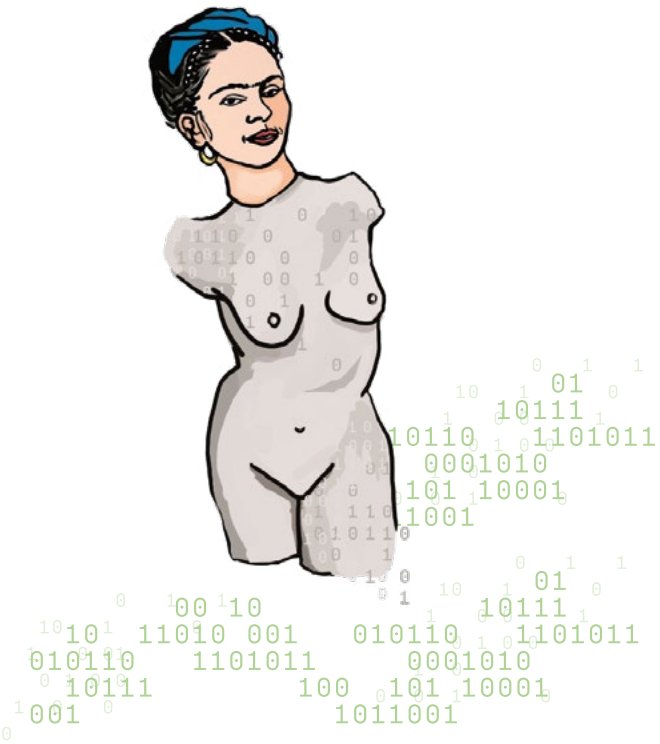
Lernziele
- Die SuS erkennen, dass KI-manipulierte bzw. KI-generierte Medieninhalte (Text, Bild, Video und Audio) missbräuchlich genutzt werden können.
- Sie lernen verschiedene Missbrauchsszenarien kennen und können diese benennen.

Unterrichtsstunden á 45 Minuten
1–2

Methoden und Material
- Multimediale Impulse anhand eines YouTube-Videos („Historische Bildmanipulation – Photoshop der Geschichte“) bzw. einer Bildreihe („Klassiker der Bildmanipulation“)
- Kooperative Gruppenarbeit mit dem AB „Deep Fake – Deep Impact“
- Recherchieren und Entdecken mithilfe einer Linksammlung

Zugang zu Internet/PC
ja

Klassenstufe
Empfohlen ab 7./8. Klasse



Einstieg

Die Manipulation von Bildern, Videos oder Audiodateien ist kein neues Phänomen. Insbesondere im politischen Kontext wird dieses Mittel häufig eingesetzt, um Einfluss auf Meinungen und Stimmungen in der Gesellschaft zu nehmen. Mit der **Digitalisierung** und der **künstlichen Intelligenz** haben sich die Möglichkeiten der Manipulation jedoch grundlegend verändert. Besonders die Fortschritte im Bereich der **generativen KI** haben bisherige Grenzen verschoben und neue Wege für digitale Fälschungen eröffnet. Sowohl in Bezug auf den Zugang, die Kosten, die Zeit, die Menge sowie die Wirkmacht durch qualitativ hochwertigere Fakes.

Zeigen Sie zum Einstieg das **Video** von MrWissen2go mit dem Titel **„Historische Bildmanipulationen – Photoshop der Geschichte“**
→ www.youtube.com/watch?v=RQdEKiWqHFse

Einstiegs-Alternative:

Zeigen Sie den SuS zum Einstieg Beispiele aus diesen **„Klassikern“ der Bildmanipulation**. Die SuS sollen bei jedem der Vorher-/Nachher-Beispiele zeigen, welche Unterschiede es zwischen den Original-Bildern und den Manipulationen gibt.
→ www.spiegel.de/fotostrecke/manipulierte-bilder-fotostrecke-107186.html

Fragen an die SuS:

- Aus welchen Gründen wurden Fotos, schon lange bevor es Computer und das Internet gab, manipuliert?
- Wenn ihr die beiden Bilder (Original – Manipulation) aus den „Klassikern“ der Bildmanipulation miteinander vergleicht: Was hat sich durch die Manipulation verändert? Welcher Eindruck wird nun erweckt?
- Kennt ihr andere bekannte Bild- oder Video-Manipulationen?

Die Macht der Bilder

Man kennt Redewendungen wie „Ein Bild sagt mehr als tausend Worte“ oder „Bilder bleiben im Kopf“ oder „Ich glaube es erst, wenn ich es sehe!“. Manche behaupten auch, der Mensch sei ein „Augentier“. In allen Aussagen steckt ein wahrer Kern. Denn der Sehsinn ist – zumindest in unserem Kulturkreis – einer unserer zentralsten Sinnesorgane über den wir unsere Umgebung wahrnehmen. Unser Gehirn reagiert auf Bilder sogar um ein Vielfaches schneller als auf Texte. Damit sind **Bilder** für uns ein **wichtiger Informationsträger**.

Idee: Besprechen Sie mit den SuS, warum uns gerade Bilder (und Bewegtbilder/Videos) so stark im Gedächtnis bleiben. Mehr Informationen zu „Macht der Bilder“ finden Sie im Sachteil in dem Kapitel „Desinformation und Demokratiegefährdung“ (s. Seite 25).

Erarbeitung

Aufgabe: Deep Fake – Deep Impact

Die SuS finden sich in Kleingruppen zusammen. Teilen Sie das **AB „Deep Fake – Deep Impact“** aus, dass die SuS in den jeweiligen Gruppen bearbeiten sollen. Alternativ kann das AB auch frontal mit der ganzen Klasse bearbeitet werden. Ziel des ABs ist, dass die SuS in einem ersten Schritt anhand von fiktiven Fall-Beispielen erklären können, welches Missbrauchsszenario jeweils dargestellt wird und welche Rolle Deepfakes dabei spielen. Die SuS werden auch befragt, ob sie noch weitere Missbrauchsszenarien durch Deepfakes kennen. Abschließend sollen die möglichen Folgen von Deepfakes erarbeiten werden. Dabei soll betrachtet werden, welche Auswirkungen Deepfakes auf Einzelpersonen, bestimmte Gruppen und die Gesellschaft haben können, wenn sie dazu genutzt werden, um Menschen zu täuschen bzw. ihnen zu schaden.
Geben Sie den SuS hierfür 20–25 Minuten Zeit. Bei jüngeren SuS wird empfohlen, die drei Fälle nacheinander zu bearbeiten. So können Lehrkräfte bei jedem Fall direkt auf mögliche Verständnisfragen der SuS eingehen.

Während die Gruppen das AB bearbeiten, schreiben Sie die Begriffe „Risiken“, „Gründe“ und „Folgen“ auf eine Präsentationsfläche (Tafel, Flipchart, Präsentations-PC, etc.). Dort werden nach der Gruppenarbeit die Ergebnisse der SuS gesammelt.

Hilfestellung

Sie können mit Ihren SuS auch über aktuelle **Beispiele aus der medialen Berichterstattung** diskutieren, um auf weitere Missbrauchsszenarien durch Deepfakes aufmerksam zu machen. Unter dem folgenden Link gelangen Sie zu einer Sammlung von Medienbeiträgen, in denen die Rolle von Deepfakes im Kontext verschiedener Risiken thematisiert wird (z. B. Desinformation oder Cybermobbing).
→ www.klicksafe.de/materialien/deep-fake-deep-impact

Sicherung

Fragen an die SuS:

- Was kann passieren, wenn Deepfakes für schlechte bzw. falsche Zwecke genutzt werden?
- Welche Gründe bzw. Absichten können dahinterstecken?
- Welche Folgen können damit verbunden sein?
- Welche der genannten Risiken durch Deepfakes sind für euch neu?

Auswertung der Aufgaben:

Besprechen Sie die drei Fallbeispiele aus Aufgabe 1. Notieren Sie auf der Präsentationsfläche die von den SuS dort identifizierten Risiken (Cybermobbing/Sextortion/Cybergrooming) sowie die damit verbundenen Gründe bzw. Absichten der Täter*innen und Folgen für die Betroffenen. Ergänzen Sie mithilfe des Lösungsblattes ggf. fehlende Aspekte.
Besprechen Sie mit den SuS anschließend weitere Missbrauchsszenarien durch Deepfakes (Aufgabe 2) und erklären Sie diese. Legen Sie dabei den Fokus auf folgende Risiken: (rechtsextreme) Desinformation, sexuelle Gewalt (Stichwort: Deepfake-Pornos oder Deepnudes) und Identitätsdiebstahl. Weiterführende Informationen finden Sie hierzu in dem Kapitel „Können Deepfakes gefährlich sein?“ (s. Seite 21).

DEEP FAKE – DEEP IMPACT

1

Noah wird bloßgestellt

Noah geht in die 8. Klasse, postet ab und zu etwas auf Social Media und ist in der Theater-AG. In der Schule hat er es nicht leicht: Er hat **kaum Freunde** und wird schon lange von vielen in der Klasse **ausgegrenzt**. Einige lästern immer wieder hinter seinem Rücken, machen sich über seine Theaterauftritte lustig und teilen gemeine Memes über ihn auf WhatsApp. Noah fühlt sich immer schlechter und würde am liebsten gar nicht mehr in die Schule gehen. Vor allem seit vorgestern plötzlich dieses **Video auf TikTok** aufgetaucht ist. Es zeigt angeblich Noah, wie er in der Schultoilette vor dem Spiegel laut Selbstgespräche führt. Noah ist total geschockt: Die Stimme im Video klingt genau wie seine und auch das Gesicht sieht aus wie er. Die ganze Schule kennt das Video und **viele glauben, dass es echt ist**. Online werden richtig verletzende Dinge kommentiert – für Noah ist es kaum auszuhalten. Sein Kopf dreht sich und er fühlt sich völlig allein.

Aufgabe 1:
Wenn Deepfakes zum Problem werden …

Lest euch die drei fiktiven Fälle durch. Die Situationen sind erfunden, beruhen aber auf Ereignissen, die so oder so ähnlich tatsächlich schon vorgekommen sind.

Formuliert für jedes der Beispiele:

- Was steckt dahinter? Was ist hier das Risiko? Welche Absichten werden verfolgt?
- Welche Rolle spielen dabei Deepfakes? Warum stellen Deepfakes hier ein Problem dar?

Deepfakes sind Texte, Bilder, Videos oder Audios, die mithilfe von künstlicher Intelligenz verändert oder sogar komplett neu erschaffen wurden. Das Besondere: Diese Inhalte können täuschend echt wirken. Es kann zum Beispiel so aussehen, als ob echte Menschen Dinge tun oder sagen, die in Wirklichkeit (so) nie passiert sind. Deepfakes können Spaß machen, aber auch zu Problemen führen bzw. gefährlich sein, weil sie zu Missverständnissen oder sogar Lügen führen können.

2

Leila und der falsche „Freund“

Leila ist 13, liebt Musik und teilt ihre Leidenschaft auch auf Instagram. Eines Tages bekommt sie eine **Freundschaftsanfrage von Finn**. In seinem Profil sieht Leila, dass er 15 ist, und ähnliche Interessen hat. Er wirkt echt süß und Leila nimmt die Anfrage an. Seit ein paar Tagen schreiben die beiden auf WhatsApp und schicken sich Selfies und Sprachnachrichten hin und her. Leila mag seine Stimme, er klingt supernett. Finn macht ihr **oft Komplimente** über ihren Musikgeschmack und ihr Aussehen. Nach einer Weile fängt Finn an, **immer persönlichere** Fragen zu stellen – zum Beispiel wo sie wohnt, in welche Schule sie geht und ob sie ihm ein Ganzkörper-Selfie senden kann. Leila geht auf alles ein – sie mag ihn und denkt sich nichts dabei. Neuerdings wird Finn direkter, wenn er schreibt. Er fragt sie nach ihren sexuellen Erfahrungen und ob sie mal **Bilder in Unterwäsche** schicken kann. Leila fängt an, sich etwas unwohl zu fühlen und weiß nicht, wie sie reagieren soll. Als sie ausweichend antwortet, fängt Finn an, sie **immer mehr unter Druck zu setzen**.

3

Luca wird erpresst

Luca ist 15, liebt Breakdance und postet viele Tanz-Videos von sich. Eines Tages bekommt er eine **Nachricht von einem Mädchen** namens „Marie“. Sie ist 15, folgt ähnlichen Seiten wie Luca und schreibt, dass er richtig gut tanzen kann. Luca findet sie interessant und fängt an, mit ihr zu schreiben – erst auf TikTok, später dann auf Whats-App. Sie schicken sich Selfies hin und her und Marie **flirtet immer mehr** mit ihm. Luca ist happy und hofft, sie bald mal zu treffen. Zwei Wochen später sendet Marie ihm **plötzlich** ein **Video**. Darin ist ein junger Mann zu sehen, der sich selbst befriedigt. Luca ist völlig geschockt, denn das **Gesicht sieht aus wie seins**! In der Nachricht **fordert** Marie **250 €** und droht das Video an seine Kontakte weiterzuleiten, wenn er nicht zahlt. Luca ist total verzweifelt und hat Angst, dass sie das tatsächlich macht.

DEEP FAKE – DEEP IMPACT

Aufgabe 2:
Weitere Fälle von problematischen Deepfakes

Fallen euch noch andere Situationen oder prominente Beispiele ein, in denen Deepfakes eingesetzt werden, um Menschen zu täuschen oder ihnen zu schaden? Warum werden solche Fakes erstellt? Welche Gründe könnten dahinterstecken?

Erstellt hier eine Sammlung von Beispielen.

Aufgabe 3:
Folgen von problematischen Deepfakes

Welche Folgen kann es haben, wenn Deepfakes für irreführende und falsche Zwecke erstellt und verbreitet werden? Schreibt auf, welche Probleme dadurch für einzelne Personen, bestimmte Personengruppen und die ganze Gesellschaft entstehen können. Beachtet dabei auch die Fälle aus Aufgabe 1.

Sammelt mögliche Folgen und notiert sie hier.

Probleme für einzelne Person:

Probleme für bestimmte Menschengruppen:

Probleme für ganze Gesellschaft:

Auflösung

Aufgabe 1

Fall 1: Noah wird bloßgestellt

Was steckt dahinter? Bei dem fiktiven Beispiel handelt es sich um einen Fall von **Cybermobbing**. Von „Cybermobbing“ spricht man, wenn jemand wiederholt und absichtlich über digitale Medien beleidigt, bloßgestellt oder bedroht wird. Täter*innen verbreiten verletzende Kommentare, peinliche Aufnahmen oder Lügen über Online-Kanäle wie WhatsApp, TikTok, Instagram oder Online-Spiele. Solche Angriffe können rund um die Uhr stattfinden, sich online schnell verbreiten und hohe Reichweite erlangen. Cybermobbing ist eine Form von psychischer Gewalt und kann zu schweren körperlichen und mentalen Belastungen bei den Betroffenen führen.

Rolle von Deepfakes: Angreifer*innen können Deepfakes gezielt einsetzen, um **Betroffene lächerlich zu machen**, **bloßzustellen und zu verletzen**. Denn mithilfe von KI können Bilder, Videos und Audios nach eigenen Wünschen verändert bzw. komplett neu erstellt werden. So können Betroffene zum Beispiel nackt oder bei einer sexuellen Handlung gezeigt werden (Stichwort: Deepfake-Pornos oder Deepnudes). Oder in einer illegalen, gewalttätigen oder stark emotionalen Situation. Es ist auch möglich, der Person Aussagen in den Mund zu legen, die sie nie gemacht hat. All dies kann täuschend echt wirken und Betroffene schwer belasten. Selbst scheinbar „harmlose“ und „lustige“ Szenen können für Betroffene extrem unangenehm und herabwürdigend sein, so dass sie sich (noch weiter) ausgegrenzt fühlen.

Fall 2: Leila und der falsche „Freund“

Was steckt dahinter? Bei dem fiktiven Beispiel handelt es sich um einen Fall von **Cybergrooming**. Mit „Cybergrooming“ ist gemeint, dass sich Personen im Internet gezielt an Minderjährige heranmachen, um zum Beispiel an intime Fotos zu kommen oder reale Treffen anzubahnen. Cybergrooming ist eine Form des sexuellen Missbrauchs und ist strafbar. Schon der Versuch gegenüber Kindern unter 14 Jahren kann zu einer Gefängnisstrafe von bis zu 5 Jahren führen. Täter*innen gehen dabei meist nach bestimmten Strategien vor, um das Vertrauen der Kinder und Jugendlichen zu gewinnen und auszunutzen.

Rolle von Deepfakes: Täter*innen können Deepfakes nutzen, um ihre **Identität zu fälschen**. Sie können zum Beispiel Bilder oder Videos erstellen, in denen sie jünger wirken oder wie eine ganz andere Person aussehen. Auch Sprachnachrichten können mithilfe von KI „verjüngt“ werden, so dass eine erwachsene Stimme wie die eines Jugendlichen bzw. Kindes klingt. Mithilfe von Face-Swap ist es auch möglich, Videoanrufe in Echtzeit zu manipulieren, so dass Betroffene den Eindruck haben können, mit einer Person ihres Alters zu sprechen.

Des Weiteren können Täter*innen **KI-generierte Intimaufnahmen** als Druckmittel einsetzen, um Kinder und Jugendliche damit zu erpressen (sogenanntes **Sextortion**). Zum Beispiel, um von Ihnen Geld oder (mehr) Bilder zu verlangen oder sie zu sexuellen Handlungen zu zwingen. KI-generierte Nacktbilder (Deepnudes) lassen sich bereits aus harmlosen Social Media-Fotos der Betroffenen erstellen oder – wie in diesem Fall – anhand von Aufnahmen, die den Täter*innen über WhatsApp und Co. übermittelt wurden.

Fall 3: Luca wird erpresst

Was steckt dahinter? Bei dem fiktiven Beispiel handelt es sich um einen Fall von **Sextortion**. Das ist eine Form der digitalen Erpressung (englisch „Extortion“), bei der Täter*innen Betroffene mit intimen Bildern oder Videos unter Druck setzen. Meistens wollen die Täter*innen Geld von den Betroffenen. Manchmal verlangen sie (noch mehr) Nacktbilder oder sogar ein Treffen, weil sie sexuelle Absichten haben (vgl. Cybergrooming). Es können aber auch emotionale Gründe dahinterstecken, wenn Täter*innen eine beendete Beziehung nicht loslassen können.

Rolle von Deepfakes: Bei Sextortion nutzen Täter*innen oft **falsche Identitäten**, um Vertrauen zur Zielperson aufzubauen – zum Beispiel, indem sie sich als gleichaltrige Jugendliche ausgeben. Mithilfe von Deepfakes-Tools können sie täuschend echte Bilder, Videos und Tonaufnahmen künstlich erzeugen. So lassen sich gefälschte Social Media-Profile erstellen oder sogar Videoanrufe in Echtzeit manipulieren. Um jemanden zu erpressen, **brauchen** Täter*innen zudem **keine echten Intimaufnahmen mehr**. Mithilfe spezieller KI-Tools können sie die Gesichter von echten Personen in bereits bestehende Pornofilme „hineinmontieren“ (Face Swap). Dafür reichen oft schon normale Bilder aus dem Internet.

Damit der **Fake-Porno** echt wirkt, passt die KI Bild für Bild die Bewegungen von Kopf, Lippen und Mimik an. Täter*innen können auch **Deepnudes** erstellen – das sind KI-generierte Nacktbilder, bei denen Personen auf ganz normalen Fotos digital „ausgezogen“ werden. Das Gesicht der Betroffenen bleibt unverändert, aber der nackte Körper ist „fake“.

Aufgabe 2

Mögliche Missbrauchsszenarien und Risiken durch Deepfakes:

- aus Aufgabe 1: Cybermobbing, Cybergrooming, Sextortion
- weitere Risiken: (politische) Desinformation, sexuelle Gewalt durch Bilder (Stichwort: Deepfake-Pornos oder Deepnudes), Hassrede, Identitätsdiebstahl

Mehr Infos

Weitere Informationen zu den Risiken, bei denen Deepfakes eine Rolle spielen können, finden Sie im Kapitel „Können Deepfakes gefährlich sein?“ (s. Seite 21) sowie in den zahlreichen klicksafe-Materialien und auf der Webseite www.klicksafe.de. Hier ein Auszug:

Für Lehrkräfte

- **Themenbereich klicksafe-Webseite „Cybermobbing“**
  → www.klicksafe.de/cybermobbing
- **Themenbereich klicksafe-Webseite „Sexualisierte Gewalt durch Bilder“**
  → www.klicksafe.de/sexualisierte-gewalt-durch-bilder
- **Themenbereich klicksafe-Webseite „Cybergrooming“**
  → www.klicksafe.de/cybergrooming

Für Jugendliche

- **Plakat „Achtung Deepfakes“**
  Download und Bestellung: → www.klicksafe.de/printmaterialien/achtung-deepfakes
- **Quiz „Deepfakes und Co.“**
  Zum Quiz: → www.klicksafe.de/materialien/quiz-zum-thema-deepfakes
  Das Quiz richtet sich an Jugendliche zwischen 10 und 14 Jahren (5. bis 7. Klasse).
- **„Safe News statt Fake News“ (Plakat und Quiz)**
  Plakat Download und Bestellung: → www.klicksafe.de/materialien/safe-news-statt-fake-news
  Zum Quiz: → www.klicksafe.de/materialien/quiz-zum-thema-safe-news
  Das Quiz richtet sich an Jugendliche zwischen 12 und 16 Jahren (7. bis 10. Klasse).
- **„Aktiv werden gegen Hate Speech“ (Flyer)**
  Download und Bestellung: → www.klicksafe.de/materialien/aktiv-werden-gegen-hate-speech
- **„F***, ich werde mit Nacktbildern erpresst! – So schützt Du Dich vor Sextortion“ (Flyer)**
  Download und Bestellung: → www.klicksafe.de/materialien/f-ich-werde-mit-nacktbildern-erpresst-so-schuetzt-du-dich-vor-sextortion
- **„Wehr Dich! Gegen sexualisierte Gewalt im Netz“ (Infobroschüre, Plakat und Erklärfilme)**
  Infobroschüre Download und Bestellung: → www.klicksafe.de/materialien/wehr-dich-gegen-sexualisierte-gewalt-im-netz
  Plakat Download und Bestellung: → www.klicksafe.de/materialien/wehr-dich-gegen-sexualisierte-gewalt-im-netz-warnsignale-im-chat
  Erklärfilme: → www.klicksafe.de/materialien/wehr-dich-gegen-sexualisierte-gewalt-im-netz-1-warnsignale-erkennen

Projekt 3

Jetzt wird's deep - Deepfakes selbst erstellen

Kurzbeschreibung	Die SuS sollen einen eigenen Deepfake erstellen. Anhand vorgegebener Kriterien sollen verschiedene KI-Tools miteinander verglichen werden.
Lernziele	▪ Die SuS können ein eigenes Bild mithilfe eines KI-Generators erstellen. ▪ Die SuS lernen die Grundlagen des Promptens kennen und können gezielt Prompts formulieren. ▪ Die SuS lernen, Deepfake-Tools kreativ und verantwortungsvoll zu nutzen.
Unterrichtsstunden á 45 Minuten	2–3
Methoden und Material	▪ Aktivierender Einstieg über Bilder-Rätsel ▪ Erkundung und Anwendung von KI-Tools zur Bildgenerierung (z. B. über OpenAI (DALL-E oder GPT 4o), Fobizz, SchulKI oder paddy)* ▪ Experimentieren mit Prompts und Vertiefung durch KI-gestütztes „Superprompting" ▪ Kreative Gruppenarbeit mit dem AB „Jetzt wird's deep" zur Erstellung einer Mini-Kampagne
Zugang zu Internet/PC	ja
Klassenstufe	Empfohlen ab 7./8. Klasse

*Wichtiger Hinweis

Prüfen Sie, wie Ihre Schule den Umgang mit generativen KI-Tools regelt: Gibt es datenschutzkonforme Zugänge für die SuS? Oder verwenden Sie als Lehrkraft einen eigenen Account zur frontalen Demonstration im Unterricht?
Beachten Sie dabei stets die datenschutzrechtlichen Vorgaben. **Informieren Sie sich bitte vorab über die geltenden Datenschutzrichtlinien Ihrer Schule oder** – falls dort keine konkreten Regelungen vorliegen – **über die Vorgaben ihres Bundeslandes**. Achten Sie darauf, dass diese Vorgaben bei der Nutzung von KI-Tools stets eingehalten werden.

Achten Sie als Lehrkraft vor allem darauf …

- dass keine persönlichen Daten von SuS oder anderen Personen in die KI-Systeme eingegeben werden.
- dass kein Material verwendet wird, das ggf. Urheberrechte verletzen könnte.
- dass keine Inhalte erstellt und genutzt werden, die irreführend, diskriminierend oder potenziell schädlich sein könnten.

Einstieg

Steigen Sie in die Stunde ein, indem Sie Beispiele zeigen, wie sich KI-generierte Bilder kreativ und spielerisch nutzen lassen. Bereiten Sie hierfür Beispiele vor und fordern Sie die SuS auf, herauszufinden, welches typische Sprichwort, Kompositum oder Homonym sich dahinter verbirgt. Ein Kompositum ist ein aus mehreren Wörtern zusammengesetztes Wort, wie zum Beispiel „Pfanne_kuchen". Mithilfe von generativer KI können beide Wortteile in einem Bild dargestellt werden. Ein Homonym ist ein Wort, das mehrere Bedeutungen hat, wie zum Beispiel „Bank" (Sitzbank und Geldinstitut). Mit KI können Bilder zu beiden Wortbedeutungen erstellt werden, damit die SuS erraten, welches Homonym dahintersteckt.

Idee: Sie können Gruppen bilden und diese gegeneinander antreten lassen. Wer errät schneller was zu sehen ist?

Hier finden Sie weitere Beispiele, wie sich KI-Bilder didaktisch sinnvoll nutzen lassen:
→ **joschafalck.de/ki-bilder**

Beispiele, um **Redewendungen** zu visualisieren:

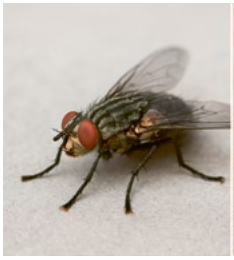
- etwas unter den Teppich kehren
- Tomaten auf den Augen haben
- gegen Windmühlen kämpfen
- der Wink mit dem Zaunpfahl
- Nägel mit Köpfen machen
- ich glaub', mein Schwein pfeift
- es sieht aus wie Kraut und Rüben
- einen Frosch im Hals haben
- das Leben ist kein Ponyhof

Beispiele, um **Komposita** zu visualisieren:

- Marmorkuchen
- Tischdecke
- Regenwurm
- Schreibtisch
- Filzlaus
- Erdbeere
- Sonnenbarsch
- Bierbauch
- Wurstwasser

Beispiele, um **Homonyme** zu visualisieren:

- Ball (Tanzball, Spielball)
- Schloss (Märchenschloss, Türschloss)
- Bienenstich (Kuchen, Insektenstich)
- Kiefer (Baum, Kopfgelenk)
- Eselsohr (Tierohr, Knick in der Buchseite)
- Leiter (Klettergerät, Führungsperson)
- Ton (Musik, Material)
- Riegel (Schokoriegel, Türriegel)
- Mutter (Mama, Schraubenmutter)



Idee

Lassen Sie die SuS mithilfe eines Bildgenerators ein Memory-Spiel erstellen, bei dem sie für jedes Homonym zwei verschiedene Bedeutungen visualisieren lassen.
Homonym-Sammlungen:
→ **www.kribbelbunt.de/artikel/news/teekesselchen?utm_source=chatgpt.com**
→ **https://malvorlagen-seite.de/teekesselchen/?utm_source=chatgpt.com**
→ **www.deutschmeisterei.de/homonym_total/?utm_source=chatgpt.com**

Aufgabe: „Prompting-Skills“

Besprechen Sie mit den SuS, **was beim Prompten zu beachten ist** und dass die **datenschutzrechtlichen Vorgaben stets einzuhalten** sind. Die SuS finden sich in Kleingruppen zusammen und erhalten am Computer Zugang zu einem DSGVO-konformen KI-Tool, um Bilder zu generieren. Bitte beachten Sie hierzu den Infokasten „Wichtiger Hinweis“ (s. Seite 46). Die Tipps für erfolgreiches Prompten finden Sie auf dem AB „Jetzt wird’s deep – Deepfakes selbst erstellen“.
Die SuS erhalten anschließend den Auftrag, sich eigene Beispiele für Redewendungen, Komposita oder Homonyme zu überlegen, die sie mithilfe eines KI-Tools visuell generieren sollen. Alternativ können Sie den SuS Beispiele vorgeben. Planen Sie hierfür ca. 10–15 Minuten Zeit ein.
Jede Gruppe stellt der Klasse ihr bestes Ergebnis bzw. ihr Lieblingsbild vor. Die anderen Gruppen sollen erraten, welches Sprichwort, Kompositum oder Homonym sich dahinter verbirgt.

Fragen an die SuS:

- Was hat bei der Umsetzung gut funktioniert?
- Wo hat die KI noch Probleme bei der Umsetzung?

Verbessert euren Prompt mithilfe der KI!

Idee: Jede Gruppe versucht nun mithilfe eines Chatbots den Prompt ihres Favoriten zu verbessern. Ziel ist es, einen besonders guten Prompt zu erstellen – einen sogenannten **Superprompt** (siehe Infokasten mit den Tipps zum Prompten). Die SuS können so erkennen, wie stark der Prompt das Ergebnis beeinflussen kann. Nachdem die Gruppen ihren Superprompt erstellt haben, geben sie diesen in den Bildgenerator ein und erstellen ein neues Bild ihres Favoriten. Vergleichen und besprechen sie gemeinsam die Ergebnisse (Bild alter Prompt vs. Bild neuer Prompt).

Fragen an die SuS:

- Was ist anders?
- Wie hat der neue Prompt das Ergebnis verändert?
- Ist das Bild nun besser geworden?
- Was habt ihr durch den „Superprompt“ gelernt?

Alternativ können Sie die Bildgenerierung frontal mit der ganzen Klasse durchführen, um zu zeigen, wie sich Deepfake-Bilder kreativ nutzen lassen. Nehmen Sie hierfür Redewendungen, Komposita oder Homonyme von den SuS entgegen und setzen Sie diese um. Sollten keine Beispiele genannt werden, nutzen Sie die Beispiele aus der Tabelle (siehe unten).

Erarbeitung

Aufgabe: Mini-Kampagne „KI kann gefährlich sein!“

Die SuS erstellen nun in Einzel- oder Kleingruppenarbeit eine **Mini-Kampagne über die negativen Folgen von Deepfakes**. Hierfür sollen sie ein Poster unter Verwendung von Deepfake-Tools erstellen. Die Aufgabenstellung finden die SuS auf dem Arbeitsblatt. Sollten Ihre SuS Anregungen brauchen, können Sie die Bilder der Beispiel-Kampagnen im Anhang zeigen. Ziel ist es, sich kritisch mit bildgenerierenden Deepfake-Tools auseinanderzusetzen und einen bewussten und verantwortungsvollen Umgang damit zu erlernen. Auf **negative Folgen** wie die Verbreitung von Falschinformationen von KI-generierten Inhalten wird in **Projekt 2** eingegangen. Klären Sie die SuS vorher über die DSGVO-konforme Nutzung von Bildgeneratoren auf (s. Seite 46, Kasten „Wichtiger Hinweis“).

Idee: An Projekttagen können sich die SuS noch intensiver mit der Kampagne auseinandersetzen und zum Beispiel einen Video-Beitrag, ein Impro-Theaterstück etc. konzipieren und umsetzen.

Sicherung

Die Poster der SuS können in einem Galeriegang ausgestellt oder in einer anderen Form untereinander geteilt werden. Es können drei Ergebnisse ausgesucht werden, die in der Schule Verbreitung finden sollen.

JETZT WIRD'S DEEP – DEEPFAKES SELBST ERSTELLEN

Startet eine Kampagne mit dem Titel **„KI kann gefährlich sein!“** für eure (jüngeren) Mitschüler*innen und nutzt hierfür eure Prompting-Skills! **Ziel der Kampagne** ist, eure Peers **über die Gefahren von Deepfakes aufzuklären**.

Mögliche Gefahren durch Deepfakes sind zum Beispiel:

Cybermobbing

Desinformation

Cybergrooming

Hassrede

Identitätsdiebstahl

Sextortion

Frauenhass (Misogynie)

Aufgabe:

- Erstellt ein kreatives Meme oder Poster mithilfe von Deepfakes-Tools.
- Macht mit eurem Beitrag deutlich, warum es wichtig ist, dass man Inhalten aus dem Netz kritisch begegnet!
- Zeigt, welche Folgen KI-generierte Inhalte haben können – z. B. für einzelne Personen, bestimmte Gruppen oder auch die ganze Gesellschaft!
- Überlegt euch Tipps, wie eure Mitschüler*innen Deepfakes erkennen und mit ihnen umgehen können. Das klicksafe-Plakat „Achtung Deepfakes“ kann euch dabei helfen.

Hinweise für erfolgreiches Prompten:

- Schreibt kurze, präzise Sätze, die leicht zu verstehen sind.
- Beschreibt, was ihr sehen wollt (Thema, Situation, eine bestimmte Szene oder Handlung).
- Beschreibt Details, die im Bild vorkommen sollen.
- Beschreibt die Stimmung und Emotionen im Bild mit Adjektiven.
- Gebt Anweisungen zum Stil (z. B. fotorealistisch, Comic-Stil, im Stil eines bzw. einer bestimmten Künstler*in, etc.).

Achtung Datenschutz!

- Gebt keine persönlichen Daten von euch oder anderen Personen in das KI-Tool ein.
- Verwendet kein Material, das Urheberrechte verletzen könnte.
- Erstellt und nutzt keine Inhalte, die irreführend, diskriminierend oder potenziell schädlich sein könnten.

Tipp: Mit der KI zum Superprompt – So optimiert ihr euer Prompting!
Das Ergebnis entspricht nicht euren Wünschen? Dann fragt doch einfach einen KI-Chatbot, wie ihr euren Prompt formulieren und optimieren könnt.
Mehr Informationen unter: www.manuelflick.de/blog/5-prompting-tipps

Beispiele für Arbeitsergebnisse

Ermutigen Sie Ihre SuS, eigene kreative Ideen für die Mini-Kampagne zu entwickeln. Vielleicht haben sie ganze andere, spannende Ansätze. Sollten Ihre SuS Hilfe für die Aufgabe benötigen, können Sie diese beiden Beispiel-Kampagnen als Anregung zeigen, um den Einstieg zu erleichtern. Beide Bilder wurden mithilfe von KI generiert.

Quelle: ChatGPT 4o, OpenAI (16.4.2025)

Quellenverzeichnis

Fußnoten:

[1] Transformer-Modelle spielen auch bei der Bildverarbeitung eine Rolle. Bei bildbasierten Deepfakes sorgen sie zum Beispiel dafür, dass über lange Bildsequenzen hinweg, zwischen den einzelnen Frames und Bildteilen ein sinnvoller Zusammenhang besteht.

[2] GPT-Modelle bilden auch die grundlegende Architektur des Chatbot ChatGPT. Mehr Informationen zu ChatGPT unter: www.klicksafe.de/materialien/wie-verlaesslich-ist-chatgpt.

[3] In Anlehnung an Albrecht (2023).

[4] Bei den Bots der „alten" Generation handelt es sich um textbasierte Schnittstellen, die auf Grundlage von vorher klar definierten Regeln funktionieren, also festgelegte Aktionen ausführen. Wenn man dieser Art von Bot also eine Frage stellt, so folgt das System starr einem vorgegebenen Gesprächsablauf und man erhält eine „Standard-Antwort". Die neuen Chatbots wie ChatGPT & Co. sind in der Lage, die Texteingaben der Nutzenden zu analysieren und darauf zu reagieren. So können sie dynamischer auf Anfragen eingehen und eine menschenähnliche Interaktion simulieren.

[5] Bei dem Launch der Version GPT-4o wurden einige neue Funktionen vorgestellt. Darunter auch der „Advanced Voice Mode". Dieser ermöglicht eine natürliche Unterhaltung mit dem Chatbot in Echtzeit. Der Bot kann bestimmte Emotionen der sprechenden Person erkennen und darauf passend reagieren – ähnlich wie ein Mensch das macht. Die KI-generierte Stimme variiert in ihrer Tonlage, Emotionalität und Geschwindigkeit, wodurch sie „menschlich" klingt. Vgl. ARD KI-Podcast (2024).

[6] Auch anhand von Bildern lassen sich über sogenannte Image-to-Video-Generatoren KI-generierte Videos erstellen.

[7] Vgl. Roswandowitz et al. (2024).

[8] Online unter: https://innovation.dw.com/sharpen-your-senses/index.html.

[9] Das Kapitel lehnt sich an die Publikation von Runge und Karaboga (2024) an und erweitert die genannten Anwendungsfelder um weitere Aspekte. Die Auflistung erhebt keinen Anspruch auf Vollständigkeit.

[10] In Anlehnung an die Seite der Bundeszentrale für politische Bildung, die auf die Chancen von Deepfakes für die Demokratie eingeht. Online unter: www.bpb.de/lernen/bewegt-bild-und-politische-bildung/556803/chancen-fuer-die-demokratie.

[11] Vgl. u.a. Forschungsarbeiten von Prof. Gevers der University of Amsterdam Gevers (2024) sowie der Digitalisierungsinitiative der Zürcher Hochschulen (DIZH).

[12] Die Risikokategorien lehnen sich an die Publikation von Runge und Karaboga (2024) an und erweitert die dort genannten Anwendungsfelder um weitere Aspekte. Die Auflistung erhebt keinen Anspruch auf Vollständigkeit.

[13] Um einen Deepfake zu erstellen, der der vertrauten Person zum Verwechseln ähnlich klingt, zeichnen Betrüger heimlich ihre Stimme auf. Zum Beispiel während eines Gesprächs, Telefonats oder aus öffentlich zugänglichen Quellen wie sozialen Medien. Kurze Stimmsequenzen reichen meist aus, um die Stimme zu klonen und einen sogenannten „replay-basierten" Sprach-Deepfake zu erstellen.

[14] Vgl. Institute for Strategic Dialogue (2024); Pranshu und Oremus (2024).

[15] Vgl. Prof. Dr. Lischka (2024).

[16] Vgl. Europäische Kommission: KI Gesetz. Online unter: https://digital-strategy.ec.europa.eu/de/policies/regulatory-framework-ai.

Literaturnachweise:

Adobe (o.J.): *CGI-Animation: Alles, was du über das Thema wissen musst.* Online unter: www.adobe.com/de/creativecloud/animation/discover/cgi-animation.html [06.02.2025].

Albrecht, Steffen (2023): *ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. Büro für Technikfolge-Abschätzung beim deutschen Bundestag.* Online unter: www.bundestag.de/resource/blob/944148/30b0896f6e49908155fcd01d77f57922/20-18-109-Hintergrundpapier-data.pdf [01.07.2024].

Alpuim, Margarida und Ehrenberg, Katja (2023a): *Warum Bilder so wirkmächtig sind.* In: Bonn Institut. Online unter: www.bonn-institute.org/news/psychologie-im-journalismus-5#wie-das-gehirn-bilder-verarbeitet-101269 [28.01.2025].

Alpuim, Margarida und Ehrenberg, Katja (2023b): *Verzerrtes Weltbild: Kognitive Filter und Denkfehler verstehen und vermeiden.* Online unter: www.bonn-institute.org/news/psychologie-im-journalismus-1 [28.01.2025].

ARD Brisant (2024, 17.08.): *Modelabel Mango wirbt mit KI-Models.* Online unter: www.ardmediathek.de/video/brisant/modelabel-mango-wirbt-mit-ki-models/das-erste/Y3JpZDovL21kci5kZS9iZWl0cmFnL2Ntcy9hZjAxOTc3OS03OWNlLTQ0MjAtYTNkZi0wMGJjMThlZWU0ZTk [03.02.2025].

ARD KI-Podcast (2024, 24.10.): *Chatbots mit Stimme: Wann wird KI zu menschlich?* Online unter: www.ardaudiothek.de/episode/der-ki-podcast/chatbots-mit-stimme-wann-wird-ki-zu-menschlich/ard/13833607/ [03.02.2025].

Bastian, Matthias (2021, 27.07.): *KI-Forschung. Mehr Retro Star Wars? Lucasfilm schnappt sich Deepfaker.* Online unter: https://the-decoder.de/mehr-retro-star-wars-lucasfilm-schnappt-sich-youtube-deepfaker/ [06.02.2025].

Behrmann, Inken (2024): *Kamala Harris for President: Wie frauenfeindlich wird der Wahlkampf in den USA?* Online unter: https://blog.campact.de/2024/07/kamala-harris-wahlkampf-frauenfeindlich-misogynie-sexismus/ [12.03.2025].

Böhm, Marcus (2024, 16.02.): *Text-zu-Video-Generator Sora OpenAI erzeugt nun auch KI-Videos.* In: Spiegel Online. Online unter: www.spiegel.de/netzwelt/web/sora-von-openai-text-zu-video-generator-erzeugt-ki-videos-a-63c96198-16de-41bc-af0e-e14927a8f02b [21.01.2025].

Bundesamt für Sicherheit in der Informationstechnik (o.J.): *Deepfakes – Gefahren und Gegenmßanahmen.* Online unter: www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html [13.06.2024].

Danry, Valdemar/ Leong, Joanne/ Pataranutaporn, Pat/ Tandon, Pulkit und Liu, Yimeng et al. (2022): *AI-Generated Characters: Puting Deepfakes to Good Use. Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems.* Online unter: https://dspace.mit.edu/bitstream/handle/1721.1/146283/3491101.3503736.pdf?sequence=1 [06.02.2025].

Das NETTZ17 (Hrsg.) (2024): *Lauter Hass - leiser Rückzug. Wie Hass im Netz den demokratischen Diskurs bedroht. Ergebnisse einer repräsentativen Befragung. Berlin.* Online unter: https://kompetenznetzwerk-hass-imnetz.de/download_lauterhass.php [19.02.2025].

Deutsche Welle (2024, 17.05.): *K-Pop und KI: Für den internationalen Erfolg.* Online unter: https://www.dw.com/de/k-pop-und-ki-f%C3%BCr-den-internationalen-erfolg/video-69109732 [06.02.2025].

Deutsche Welle Innovation (2023): *How to verify? Explore workflows and tools for verification.* Online unter: https://www.howtoverify.info/ [25.06.2024].

Deutsches Forschungszentrum für künstliche Intelligenz (2024, 07.06.): *Deepfakes im Visier: KI als Waffe gegen digitale Manipulation.* Online unter: www.dfki.de/web/news/mit-ki-gegen-deepfakes [25.06.2024].

Deutschlandfunk (2021, 23.10.): *Deepfakes. Wenn Künstliche Intelligenz Geschichte, Stars und Präsidenten fälscht.* Online unter: www.deutschlandfunk.de/deepfakes-wenn-kuenstliche-intelligenz-geschichte-stars-und-100.html#:~:text=Der%20Reddit%2DUser%20mit%20dem,auf%20gro%C3%9Fe%20%E2%80%9EResonanz%E2%80%9C%20stie%C3%9F [25.06.2024].

Deutschlandfunk Nova (2021, 02.03.): *Myheritage. Deepfake-Technologie lässt Vorfahren auferstehen.* Online unter: www.deutschlandfunknova.de/beitrag/myheritage-deepfake-technologie-laesst-vorfahren-wieder-auferstehen [30.01.2025].

Digitalisierungsinitiative der Zürcher Hochschulen (2024): *Deepfakes und Virtual Reality in der Psychotherapie.* Online unter: www.dizh.uzh.ch/2025/01/21/deepfakes-und-virtual-reality-in-der-psychotherapie/ [06.02.2025].

Dirik, Iskender (2018): *Deep Learning einfach gemacht. Die Komponenten eines Deep-Learning-Systems und was diese verdammten nerdigen Buzzwords eigentlich bedeuten.* Online unter: https://news.microsoft.com/uploads/prod/sites/40/2018/11/Deep-Learning-einfach-gemacht.pdf [26.06.2024].

Donath, Andreas (2024, 08.12.): *Aurora: X stellt neues KI-Tool zur Bilderzeugung vor.* In: Golem. Online unter: www.golem.de/news/aurora-x-stellt-neues-ki-tool-zur-bilderzeugung-vor-2412-191521.html [19.02.2025].

Dongus, Ariana (2024, 13.08.): *Nackte Gewalt. Künstliche Nacktbilder zu erstellen wird technisch immer einfacher. Doch wie kann man sich gegen Deepnudes wehren?* Online unter: www.fluter.de/deepfakes-deepnudes

El Ouassil, Samira (2024, 02.02.): *Pornografische Deepfakes. Die digitale Waffe gegen die Frau.* In: Spiegel Online. Online unter: www.spiegel.de/kultur/pornografische-deepfakes-digitale-waffe-gegen-die-frau-kolumne-a-42394844-47f3-4b76-9f89-1067a81e5874 [12.03.2025].

EU-Agentur für Grundrechte (2023): *Online content moderation - Current challenges in detecting hate speech.* Online unter: https://fra.europa.eu/en/news/2023/online-hate-we-need-improve-content-moderation-effectively-tackle-hate-speech [21.02.2025].

Europäische Kommission: *KI Gesetz.* Online unter: https://digital-strategy.ec.europa.eu/de/policies/regulatory-framework-ai [18.03.2025].

Friedel, Lea Joy (2025, 17.01.): *Deepfakes: Die Ohnmacht angesichts dieser Bilder.* In: Zeit Online. Online unter: www.zeit.de/kultur/2025-01/deep-fakes-ki-internet-misogynie-pornographie [19.02.2025].

Gent, Edd (2024, 06.04.): *Quelloffene KI: Warum die Techbranche darüber streitet, was das überhaupt ist.* In: Heise Online. Online unter: www.heise.de/hintergrund/Quelloffene-KI-Warum-die-Techbranche-darueber-streitet-was-das-ueberhaupt-ist-9665378.html [04.07.2024].

Gevers, Theo (2024): *Deepfakes with therapeutic effect. University of Amsterdam, Informatics Institute.* Online unter: https://ivi.uva.nl/content/news/2024/09/deepfakes-with-therapeutic-effect.html?cb [06.02.2025].

Gilbert, Max (2024, 02.08.): *Bots, Bilder, Deepfakes: Wie KI für Desinformation genutzt wird.* In: BR24. Online unter: www.br.de/nachrichten/netzwelt/bots-bilder-deepfakes-wie-ki-fuer-desinformation-genutzt-wird-faktenfuchs,UKCyvx8 [05.03.2025].

Gockel, Jola (2024): *The current state of pornographic Deepfakes. A Science and Technology Studies Perspective.* In: The Maastricht Journal of Liberal Arts, Vol. 15 (2024), S. 15-25. Online unter: https://openjournals.maastrichtuniversity.nl/MJLA/article/view/1005 [21.02.2025].

Grellmann, Martin (2021): *10 Deep Learning Algorithmen, die Sie kennen sollten.* Online unter: https://martin-grellmann.de/10-deep-learning-algorithmen-die-sie-kennen-sollten#Arten [01.07.2024].

Hains, Tim (2024, 17.06.): *Karine Jean-Pierre: „Cheap Fake" Videos Of Biden Tell You Everything You Need To Know About How Desperate Republicans Are.* Online unter: www.realclearpolitics.com/video/2024/06/17/karine_jean-pierre_cheap_fake_videos_tell_you_everything_you_need_to_know_about_how_desperate_republicans_are.html [05.03.2025].

Harwell, Drew (2019, 24.05.): *Faked Pelosi videos, slowed to make her appear drunk, spread across social media.* In: Washington Post. Online unter: www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [22.01.2025].

Hate Aid (2021): *Report „Grenzenloser Hass im Internet – Dramatische Lage in ganz Europa".* Online unter: https://hateaid.org/eu-umfrage-grenzenloser-hass-im-internet/ [19.02.2025].

Holtermann, Felix und Bomke, Luisa (2024, 10.12.): *OpenAI veröffentlicht Videogenerator „Sora" – aber nicht in Deutschland.* In: Handelsblatt Online. Online unter: www.handelsblatt.com/technik/it-internet/kuenstliche-intelligenz-openai-veroeffentlicht-videogenerator-sora-aber-nicht-in-deutschland/100093817.html [22.01.2025].

Huckebrink, Lydia (2024, 22.04.): *Musik per Mausklick. Virtuelle Pop-Stars: Wenn Künstliche Intelligenz die Charts stürmt.* In: SWR Kultur. Online unter: www.swr.de/swrkultur/musik-jazz-und-pop/ki-in-der-musik-wann-stuermen-virtuelle-popstars-die-charts-100.html [12.03.2025].

Ignor, Sarah (2025, 08.01.): *KI-Chatbot Grok zieht in Tesla-Fahrzeuge ein.* In: Computer Bild. Online unter: www.computerbild.de/artikel/cb-News-Connected-Car-Elon-Musk-KI-Chatbot-Grok-Tesla-Fahrzeuge-xAI-X-Plattform-39324729.html [19.02.2025].

JIM-Studie (2024): *Jugend, Information, Medien.* Online unter: https://mpfs.de/studie/jim-studie-2024/ [12.03.2025].

Jones, Cameron R. und Bergen, Benjamin K. (2024): *People cannot distinguish GPT-4 from a human in a Turing test.* Online unter: https://arxiv.org/abs/2405.08007 [09.07.2024].

Justus Thies et al. (2016): *Real-time Face Capture and Reenactment of RGB Videos. Online unter. www.niessnerlab.org/projects/thies2016face.html [12.03.2025].*

Karaboga, Murat et al. (2024): *Deepfakes und manipulierte Realitäten. Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz. TA-SWISS Publikationsreihe (Hrsg.): TA 81/2024. Zolllikon: vdf.* Online unter: https://zenodo.org/records/11643644 [09.07.2024].

Kastorff, Tamara/ Müller, Maren und Selva, Clieviens (2025): *Fake News oder Fakten? Wie Jugendliche ihre digitale Informationskompetenz einschätzen und welche Rolle Schulen und Lehrkräfte dabei spielen. Erkenntnisse aus PISA 2022. Münster: Waxmann Verlag.* Online unter: www.waxmann.com/index.php?eID=download&buchnr=4993 [12.03.2025].

Kettmann, Otto (2017): *Emotionen und das Ringen um Aufmerksamkeit Der mediale Trend zu emotionsfokussierten Stilmitteln.* In: Communicatio Socialis Jahrgang 50 (2017), S. 360-368. Online unter: www.nomos-elibrary.de/10.5771/0010-3497-2017-3-360.pdf?download_full_pdf=1 [06.02.2025].

Kramer, André (2024, 03.06.): *Sieben bekannte KI-Bildgeneratoren im Test.* In: Heise Online. Online unter: www.heise.de/tests/Sieben-bekannte-KI-Bildgeneratoren-im-Test-9690639.html [01.07.2024].

Kwon, Jake und Watson, Ivan (2023, 04.10.): *„The only thing we can't do is sign autographs": The rise of virtual K-pop bands.* In: CNN. Online unter: https://edition.cnn.com/style/kpop-virtual-bands-ai-intl-hnk/index.html [03.02.2025].

Linden, Michael (2024, 20.12.): *Tencent Hunyuan Video: Kostenloses KI-Modell für Videogenerierung vorgestellt.* In: Golem. Online unter: www.golem.de/news/tencent-hunyuan-video-kostenloses-ki-modell-fuer-videogenerierung-vorgestellt-2412-191918.html [21.01.2025].

Lundberg, Ebba und Mozelius, Peter (2024): *The potential effects of deepfakes on news media and entertainment.* In: AI & Society. Journal of Knowledge, Culture and Communication. Online unter: https://link.springer.com/article/10.1007/s00146-024-02072-1#Sec14 [04.02.2025].

Magazin für Kommunikation (2013): *Menschen sind Augentiere.* In: Magazin für Kommunikation. Online unter: www.kom.de/medien/menschen-sind-augentiere/ [25.06.2024].

Masood, Momina et al. (2021): *Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward.* Online unter: www.researchgate.net/publication/349703826_Deepfakes_generation_and_detection_state-of-the-art_open_challenges_countermeasures_and_way_forward [18.07.2024].

MDR (2022, 02.05.): *Nachrichten vom Avatar | MDR.* Online unter: www.youtube.com/watch?v=zOfZlkjudS8 [03.02.2025].

MDR exactly (2023): *Missbrauch mit KI: So gefährlich sind Deepfakes.* Online unter: www.ardmediathek.de/video/exactly/missbrauch-mit-ki-so-gefaehrlich-sind-deepfakes/mdr-fernsehen/Y3JpZDovL21k-ci5kZS9zZW5kdW5nLzI4MjA0MS8yMDIzMDkxODA4MDAv-bWRycGx1cy1zZW5kdW5nLTc0MDQ [01.07.2024].

Müller, Nicolas/Pizzi, Karla und Williams, Jennifer (2024): *Human Perception of Audio Deepfakes. Cornell University.* Online unter: https://arxiv.org/abs/2107.09667 [15.01.2025].

Museum für Kommunikation (2007): *Lehrmaterial „Bilder, die lügen".* Online unter: www.mfk.ch/lehrmittel/lehrmittel-bilder-die-luegen [12.03.2025].

Museumspädagogisches Zentrum (2021): *Caesar und die Macht der Bilder.* Online unter: www.youtube.com/watch?v=Rrino-PuVhms[28.01.2025].

Nickel, Oliver (2024, 29.07.): *Elon Musks KI sammelt die Daten der User auf X.* In: Golem. Online unter: www.golem.de/news/xai-grok-x-user-erfahren-erst-jetzt-dass-ki-an-ihren-daten-trainiert-2407-187513.html [19.02.2025].

Onlinesicherheit.at (2021): *Die Entstehung von Deep Fakes auf Reddit und ihre Verbreitung.* Online unter: www.onlinesicherheit.gv.at/Services/News/Die-Entstehung-von-Deep-Fakes-auf-Reddit-und-ihre-Verbreitung.html [25.06.2024].

Paris, Britt und Donovan, Joan (2019): *Deepfakes and Cheap Fakes. The Manipulation of Audio and Visual Evidence.* Online unter: https://datasociety.net/library/deepfakes-and-cheap-fakes/ [12.03.2025].

re:publica 2024*: Deepfakes - Our New Reality? Vera Schmitt und Tim Polzehl.* Online unter: www.youtube.com/watch?v=vdu01kYlHgk [25.06.2024].

Roswandowitz, Claudia et al. (2024): *Cortical-striatal brain network distinguishes deepfake from real speaker identity.* In: Communications Biology, volume 7/711 (2024), S. 1-14. Online unter: s42003-024-06372-6.pdf [22.01.2025].

Runge, Greta und Karaboga, Murat (2024): *Deepfakes als kulturelle Praxis und gesellschaftliche Herausforderung: Zu Potentialen und Wirkungsweisen der Technologie.* In: Prof. Dr. Decker, Michael et al. (Hrsg.): Gestreamt, gelikt, flüchtig – schöne neue Kulturwelt? Digitalisierung und Kultur im Licht der Technikfolgenabschätzung. Nomos Verlag: Baden Baden, S. 329-346. Online unter: www.researchgate.net/publication/382571640_Deepfakes_als_kulturelle_Praxis_und_gesellschaftliche_Herausforderung_Zu_Potentialen_und_Wirkungsweisen_der_Technologie [29.01.2025].

Schiff, Kaylin/ Schiff, Daniel und Bueno, Natália (2024): *The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability? Cambridge University Press.* In: American Political Science Review, Vol. 119/1 (2025), S. 71-90. Online unter: https://www.cambridge.org/core/journals/american-political-science-review/article/liars-dividend-can-politicians-claim-misinformation-to-evade-accountability/687FEE54DBD7ED0C96D72B26606AA073 [12.03.2025].

Schneider, Jan (2023, 23.09.): *Deepnudes in Spanien: Schülerinnen mit KI-Nacktbildern gemobbt.* In: ZDF heute. Online unter: https://www.zdf.de/nachrichten/panorama/spanien-schuelerinnen-deepnudes-nacktbilder-100.html [12.03.2025].

The Dalí Museum (2019): *Behind the Scenes: Dalí Lives.* Online unter: www.youtube.com/watch?v=BIDaxl4xqJ4 [12.03.2025].

Vincent, James (2021, 05.03.): *Tom Cruise deepfake creator says public shouldn't be worried about 'one-click fakes'.* In: The Verge. Online unter: www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher [01.07.2024].

Wiggers, Kyle (2024, 26.11.): *Der Sora-Videogenerator von OpenAI scheint durchgesickert zu sein. Tech Crunch:* Online unter: https://techcrunch.com/2024/11/26/artists-appears-to-have-leaked-access-to-openais-sora/?guccounter=1 [21.01.2025].

WITNESS (2020): *Identity protection with deepfakes: „Welcome to Chechnya" director David France.* Online unter: www.youtube.com/watch?v=2du6dVL3Nuc [12.03.2025].

Wolter, Nadine (2024): *Mango wirbt mit KI-Motiven. Diese junge Frau gibt es nicht.* In: Spiegel Online. Online unter: www.spiegel.de/netzwelt/mango-wirbt-mit-ki-motiven-in-kampagne-fuer-sunset-dream-kollektion-a-39785de4-61a0-452d-9629-287791abb93d [12.03.2025].

X (o.J.): *Hilfe Center.* Online unter: https://help.x.com/de/using-x/about-grok [19.02.2025].

Xu, Sicheng et al. (2024): *VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time.* Online unter: www.microsoft.com/en-us/research/project/vasa-1/ [16.07.2024].

Youngs, Ian (2024, 01.05.): *FKA Twigs uses AI to create deepfake of herself.* In: BBC Online. Online unter: www.bbc.com/news/articles/c6py33gxk74o [03.02.2025].

Zalando (2024, 17.10.): *Zalando erweitert virtuelle Umkleidekabine um 3D-Avatar mit individuellen Körpermaßen der Kund*innen.* Online unter: https://corporate.zalando.com/de/technologie/zalando-erweitert-virtuelle-umkleidekabine-um-3d-avatar-mit-individuellen-koerpermassen [12.03.2025].

Hilfe- und Beratung

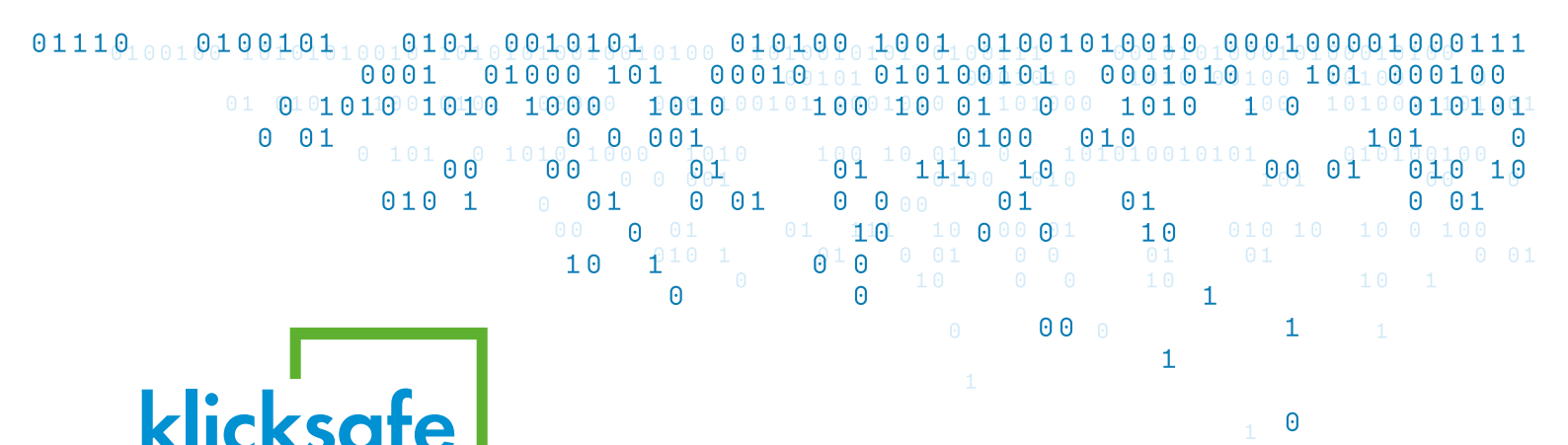
Nummer gegen Kummer:
Kinder- und Jugendtelefon (anonym und kostenlos):
116 111 (Mo–Sa 14–20 Uhr)
Elterntelefon (anonym und kostenlos):
0800 111 0 550 (Mo–Fr 9–17 Uhr | Di–Do 9–19 Uhr)
Online-Beratung per Mail oder Chat:
www.nummergegenkummer.de

JUUUPORT:
Kostenlose Beratungsstelle von jungen Menschen für junge Menschen. Online-Beratung unter:
www.juuuport.de/hilfe/beratung

HateAid:
Kostenlose Beratungsstelle für Betroffene digitaler Gewalt.
Telefon: 030 25208838 (Mo 10–13 Uhr | Do 15–18 Uhr)
Chat (Mi 15–18 Uhr | Fr 11–14 Uhr)
E-Mail: beratung@hateaid.org
www.hateaid.org/betroffenenberatung

Hilfe-Telefon sexueller Missbrauch:
Anonyme, kostenlose und mehrsprachige Hilfe und Beratung.
Telefon: 0800 22 55 530 (Mo, Mi, Fr 9–14 Uhr | Di, Do 15–20 Uhr)
Online-Beratung: https://schreib-ollie.de
www.hilfe-portal-missbrauch.de

Bundesverband Frauenberatungsstellen und Frauennotrufe:
Informationen über verschiedene Formen von digitaler Gewalt gegen Frauen sowie hilfreiche Tipps unter:
www.aktiv-gegen-digitale-gewalt.de



klicksafe

Bezugsadresse:

EU-Initiative klicksafe
Medienanstalt Rheinland-Pfalz
Turmstraße 10
D-67059 Ludwigshafen

info@klicksafe.de
www.klicksafe.de

Weitere Materialien finden Sie unter:
www.klicksafe.de/materialien

klicksafe ist das deutsche Awareness Centre im Digital Europe Programm der Europäischen Union und wird von der Medienanstalt Rheinland-Pfalz verantwortet.

Medienanstalt Rheinland-Pfalz (AöR)
Vertreten durch: Dr. Marc Jan Eumann
Turmstraße 10
D-67059 Ludwigshafen
mail@medienanstalt-rlp.de
www.medienanstalt-rlp.de

Kofinanziert von der Europäischen Union

Medienanstalt Rheinland-Pfalz