AI Skills, Powered by Values

values

www.valuesai.eu

MODUL 03

Desinformation, Deepfakes & Datenmanipulation

KI-gestützte Täuschung erkennen und verhindern

Co-funded by the European Union

Mitfinanziert von der Europäischen Union. Die geäußerten Ansichten und Meinungen sind jedoch ausschließlich die der Autor*innen und spiegeln nicht unbedingt die Ansichten der Europäischen Union oder der Stiftung für die Entwicklung des Bildungssystems wider. Weder die Europäische Union noch die fördernde Einrichtung können dafür verantwortlich gemacht werden.

Inhalt

Co-funded by the European Union

Mitfinanziert von der Europäischen Union. Die geäußerten Ansichten und Meinungen sind jedoch ausschließlich die der Autor*innen und spiegeln nicht unbedingt die Ansichten der Europäischen Union oder der Stiftung für die Entwicklung des Bildungssystems wider. Weder die Europäische Union noch die fördernde Einrichtung können dafür verantwortlich gemacht werden.

01

Desinformation

(20 Min.)

Desinformation

KI-gestützte Desinformation bezeichnet absichtlich falsche oder irreführende Inhalte, die mithilfe von KI-Tools erstellt, bearbeitet oder verbreitet werden (z. B. generative Modelle, Deepfakes oder Bots), um Menschen zu täuschen, Wahrnehmungen zu beeinflussen und demokratischen Prozessen, Institutionen oder Einzelpersonen zu schaden.

AI Skills,
Powered
by Values
values

Ein Beispiel für Desinformation

Eines der bekanntesten Beispiele für KI-gestützte Desinformation ist ein gefälschtes Bild von Donald Trumps angeblicher Verhaftung. Es kursierte 2023 in sozialen Medien und auf Nachrichtenplattformen in Europa und den USA.

Das Bild zeigte Trump, wie er von Polizisten abgeführt wurde. Eine solche Verhaftung hatte jedoch nicht stattgefunden: Das Bild wurde vollständig mit einem KI-Bildgenerator erstellt und an reale Archivfotos angelehnt.

Viele teilten das Bild mit dem Hinweis, dass es gefälscht war. Trotzdem wurden einige Menschen offenbar in die Irre geführt.

Diskussionsübung

„Wir werden wissen, dass unser Desinformationsprogramm abgeschlossen ist, wenn alles, was die amerikanische Öffentlichkeit glaubt, falsch ist.“

William J. Casey, Direktor der CIA

1. Überlege, welche wirtschaftlichen und sozialen Folgen eine solche Nachricht haben kann.
2. Sind die Schäden durch solche Aktivitäten vollständig rückgängig zu machen?
3. Das Medium mit dem mit Abstand größten Desinformationsrisiko ist das Internet. Überlege, warum das so ist.
4. Führt eine Gruppenabstimmung durch: Vertraut ihr Fernsehnachrichten mehr oder Nachrichten aus dem Internet? Wie passt das Ergebnis eurer Abstimmung zum Risiko von Desinformation?

Fehlinformation, Desinformation und Manipulation sind nicht dasselbe.

Fehlinformation	Desinformation	Manipulation
Falsche oder ungenaue Informationen, die nicht unbedingt absichtlich verbreitet werden	Falsche Informationen, die absichtlich erstellt und verbreitet werden	Veränderung von Inhalten, Kontext oder Daten, um die Empfänger*innen zu beeinflussen

02

Deepfakes

(20 Min.)

Deepfakes: Warum sie schwer zu erkennen sind

- **Ein Deepfake ist ein Medieninhalt, der mithilfe künstlicher Intelligenz erstellt oder verändert wurde, sodass es so wirkt, als hätte jemand etwas gesagt oder getan, was nie passiert ist. Ausgereifte Algorithmen können echtes Filmmaterial mit computergenerierten Bildern und Audiodaten so nahtlos verbinden, dass man es durch bloßes Hinsehen oder Hinhören oft kaum erkennt.**

- **Diese Technologie hat sich sehr schnell verbessert. Sie hinterlässt nicht mehr unbedingt offensichtliche Fehler oder „Hinweise“ wie unpassende Lippenbewegungen oder verschwommene Hintergründe. Deshalb können selbst erfahrene Journalist*innen und Expert*innen für digitale Forensik Schwierigkeiten haben, gut gemachte Fälschungen zu erkennen.**

Deepfakes: Warum sie schwer zu erkennen sind

- **Deepfakes können für Humor oder Kunst genutzt werden, werden aber zunehmend eingesetzt, um Menschen zu täuschen oder zu schädigen. Sie können Reputationen beschädigen, falsche Informationen verbreiten oder bei politischen und persönlichen Angriffen instrumentalisiert werden. Gerade weil sie so realistisch wirken, brauchen wir eine vorsichtige, fragende Haltung, wenn wir online auf überraschende Inhalte stoßen.**

Möchtest du deine digitalen Detektivfähigkeiten testen?

Die Website „Detect Fakes“ des MIT Media Lab gibt dir die Möglichkeit, echte und KI-generierte Videos anzusehen und zu entscheiden, welche echt sind.

Die Forschenden haben dieses Tool entwickelt, damit Menschen durch Ausprobieren lernen. Es gibt keinen einzigen Trick, um Deepfakes zu erkennen. Wenn du jedoch auf kleine Details in Gesichtern und Beleuchtung achtest, kannst du besser einschätzen, ob etwas manipuliert wurde.

Dein Reflexions-Toolkit

Während dieser Einheit hältst du immer wieder kurz inne und denkst über das Gesehene oder Gehörte nach. Notiere deine Gedanken in einem Heft, auf dem Smartphone oder in einer Notizen-App.

Es gibt keine einzige richtige Antwort. Deine Notizen helfen dir, online Gesehenes kritisch zu hinterfragen und sicherer mit Unsicherheit umzugehen.

01

Welche Hinweise lassen dich vermuten, dass ein Video oder Foto gefälscht sein könnte?

02

Welche Gedanken oder Fragen kommen dir, besonders wenn sich etwas „nicht richtig“ anfühlt?

03

Warum könnte jemand eine Fälschung erstellen und teilen?

Was denkst du?

Wenn du verstehst, was Deepfakes sind und warum sie schwer zu erkennen sind, kannst du später in diesem Modul die Drei-Punkte-Checkliste besser anwenden.

01

Hast du schon einmal einen Deepfake gesehen oder vermutet, dass ein Video manipuliert wurde? Wie hast du dich dabei gefühlt?

02

Wenn Fälschungen so echt wirken können: Welche zusätzlichen Schritte solltest du gehen, bevor du einen überraschenden Clip glaubst oder teilst?

03

Wie könnte Deepfake-Technologie positiv genutzt werden und wie können wir Missbrauch verhindern?

Denken: Mit Unsicherheit umgehen

Diese Aktivitäten sollen dein kritisches Denken stärken und dir helfen, dich in einer digitalen Welt zu orientieren, in der Fakten und Fiktion oft ineinander übergehen.

01

Wie schnell teilst du ein interessantes oder lustiges KI-Bild oder einen Clip? Wie entscheidest du, ob du teilst?

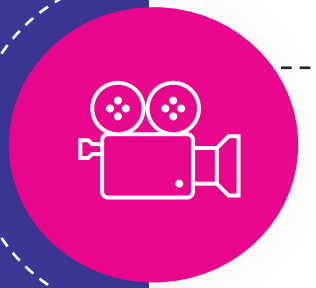
02

Was ist der Unterschied zwischen einem lustigen KI-Bild oder Clip und einem Inhalt, der Menschen schadet?

03

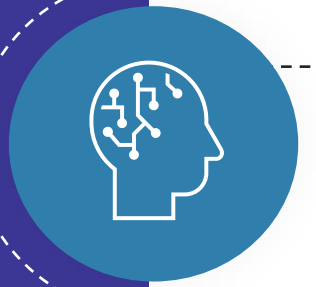
Wer könnte davon profitieren, wenn sich KI-gestützte Fehlinformation verbreitet?

Arten KI-generierter Falschinhalte – Definitionen & Beispiele



01 Fake-Videos – Videoaufnahmen, in denen eine Person scheinbar Dinge sagt oder tut, die sie nie gesagt oder getan hat; erstellt oder verändert mit KI (z. B. Deepfakes)

02 Gefälschte Kommentare/Bewertungen – KI-generierte Meinungen, Bewertungen oder Kommentare im Internet, die künstlich ein positives oder negatives Bild eines Produkts, Unternehmens oder Politikers aufbauen sollen



03 Synthetische Konten & Bots – falsche Nutzerprofile oder Bot-Konten, die von KI gesteuert werden und echte Nutzer*innen nachahmen, um die Reichweite falscher Inhalte zu erhöhen

04 Automatisch generierte Einflusskampagnen – Informationskampagnen, deren Inhalte (Texte, Grafiken, Videos, Social-Media-Posts) teilweise oder vollständig KI-generiert und für massenhafte Verbreitung automatisiert sind

Arten KI-generierter Falschinhalte – Definitionen & Beispiele

05 Fake News – Artikel oder Überschriften, die von KI erstellt oder verändert werden, damit sie wie legitime Nachrichten aussehen, tatsächlich aber falsche oder verzerrte Informationen enthalten

06 Gefälschte Zitate – Aussagen, die fälschlich Prominenten, Politiker*innen oder Wissenschaftler*innen zugeschrieben und von KI als „Zitate“ erzeugt werden

07 Fake-Bilder – Grafiken oder „Fotos“ von Ereignissen, die nie stattgefunden haben (z. B. politische Verhaftungen, Katastrophen, gesellschaftliche Ereignisse), erstellt mit KI-Bildgeneratoren

08 Gefälschte Stimmaufnahmen – synthetische Stimmen öffentlicher Personen, die für falsche Audio-Clips, Gespräche, Aussagen oder Reden genutzt werden

03

Daten- & Kontextmanipulation

(15 Min.)

Typische Datenmanipulation

- Selektive Darstellung von Daten – nur der Teil des Datensatzes wird gezeigt, der eine Behauptung stützt; der Rest wird weggelassen.
- Veränderung der Diagrammskala – eine abgeschnittene Achse oder ungleiche Abstände lassen kleine Unterschiede viel größer wirken.
- Quelle weglassen – eine Zahl, ein Diagramm oder ein Zitat wird ohne Quelle, Methodik oder Datum präsentiert.
- Falsche Statistiken – erfundene, falsche oder mathematisch fehlerhafte Zahlen werden präsentiert.

Typische Datenmanipulation

- **Aussagen aus dem Kontext reißen – ein kurzer Ausschnitt verändert die ursprüngliche Bedeutung.**
- **Wahre Daten, falsche Deutung – korrekte Zahlen führen zu einer irreführenden Schlussfolgerung.**
- **Selektiver Zeitraum – Start- und Enddatum erzeugen künstlich Wachstum, Rückgang oder Krise.**
- **Manipulative Kategorien/Stichproben – Gruppierung oder Auswahl lässt das Ergebnis überzeugender wirken.**

Praktische Beispiele

Den Trend verstecken

Nehmen wir an, wir möchten die Verkaufszahlen von vier Produkten unseres Unternehmens präsentieren. Leider verkauft sich Produkt B (rot), das ein zentraler Bestandteil unseres Angebots ist, nicht mehr so gut wie früher. Kein Problem – wir können das leicht verbergen.

Mit einem gestapelten Balkendiagramm können wir den Umsatzrückgang unauffällig machen oder zumindest weniger drastisch erscheinen lassen. Wenn wir dieselben Daten als Liniendiagramm darstellen, sehen wir die tatsächliche Lage deutlich: Der Abwärtstrend ist klar erkennbar. Kaum zu glauben, dass beide Diagramme dieselben Daten zeigen, oder?

Siehe nächste Folie

Den Trend verstecken

ICAN. (o. J.). Kurzer Leitfaden zur Manipulation in Datenvisualisierungen. Abgerufen am 17. Mai 2026 von https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16IwGIn0h

Praktische Beispiele

Skala verändern

Im ersten Diagramm beginnt die Skala bei 0. Wir sehen, dass der Unterschied, zum Beispiel zwischen Balken 2 und 7, nicht besonders groß ist.

Im zweiten Diagramm werden dieselben Daten erneut dargestellt. Hier beginnt die Skala bei 100. Dadurch wirken die Unterschiede zwischen den einzelnen Balken deutlich größer.

Siehe nächste Folie

Skalenänderung

Praktische Beispiele

Falsche Perspektive

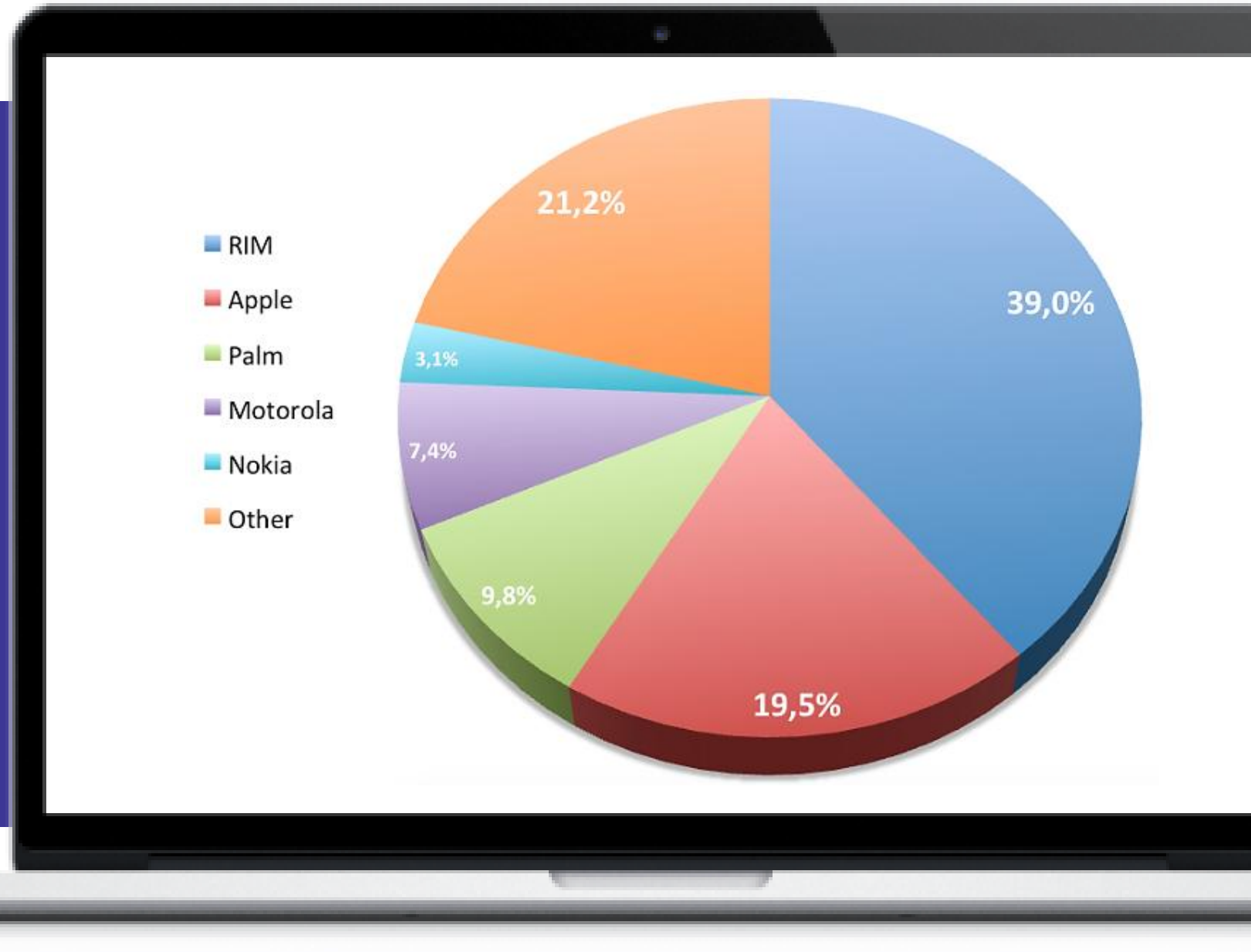
Im Jahr 2008 präsentierte Steve Jobs Apples Anteil am Smartphone-Markt. Im oberen Diagramm (eine genaue Nachbildung des von Steve Jobs gezeigten Diagramms) wirkt das Segment für Apple größer als das Segment „Andere“. Es erscheint auch nicht deutlich kleiner als der Anteil von RIM (BlackBerry). Betrachtet man jedoch die Prozentwerte, hatte Apple damals nur halb so viel Marktanteil wie RIM.

Auf den ersten Blick fällt nicht auf, dass es sich um ein 3D-Diagramm handelt. Es wurde geschickt gedreht, sodass es vertikaler wirkt. Der Effekt wird durch unterschiedliche Schriftgrößen verstärkt: Für weniger bedeutende Anteile wird eine kleinere Schrift verwendet, während RIM, Apple und „Andere“ gleich groß beschriftet sind.

Siehe nächste Folie

Falsche Perspektive

ICAN. (o. J.). Kurzer Leitfaden zur Manipulation in Datenvisualisierungen. Abgerufen am 17. Mai 2026 von https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16IwGIn0h



Praktische Beispiele

Clever gruppieren

Die Diagramme zeigen Steuerdaten aus den Vereinigten Staaten. Im linken Diagramm sehen wir das Einkommen (in absoluten Zahlen) verschiedener Gruppen nach Einkommenshöhe.

Die subtile Manipulation liegt hier in der Gruppierung der Ergebnisse in unterschiedliche Kategorien, sodass das Diagramm die gewünschte Form annimmt. So kann aus einer Darstellung, die links einer Normalverteilung ähnelt, rechts ein „Hockeyschläger“-Diagramm werden.

Praktische Beispiele

Clever gruppieren

Wenn wir das Diagramm links betrachten, kommen wir zu dem Schluss, dass Menschen aus der Mittelschicht den größten Geldbetrag besitzen. Im Diagramm rechts wurden mehrere Kategorien zu einer größeren Gruppe zusammengefasst.

Dadurch könnte man schließen, dass die obere Schicht (ab 200.000 Dollar aufwärts) den größten Geldbetrag besitzt. Dieselben Daten können zu völlig unterschiedlichen Schlussfolgerungen führen – und damit auch zu unterschiedlichen gesellschaftlichen Entscheidungen, etwa bei Wahlen.

Siehe nächste Folie

Clever gruppieren

ICAN. (o. J.). Kurzer Leitfaden zur Manipulation in Datenvisualisierungen. Abgerufen am 17. Mai 2026 von https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16IwGIn0h

04

Dein Detektiv-Toolkit

(20 Min.)

Künstliche Intelligenz kann überzeugende Fake-Medien erzeugen. Deshalb brauchst du einen strukturierten Ansatz, um zu entscheiden, ob etwas echt ist.

AI Skills,
Powered
by Values
values

Checkliste für dein Denken

Diese Punkte sind nicht narrensicher. Sie helfen dir, vor dem Teilen langsamer und genauer hinzuschauen.

1. **Details prüfen: Stimmen Finger, Spiegelungen, Licht und Schatten? Gibt es Bildfehler?**
2. **Quelle prüfen: Wer hat es gepostet? Berichten seriöse Quellen dasselbe?**
3. **Gefühl prüfen: Soll der Inhalt schockieren, wütend machen oder provozieren? Wenn ja: erst prüfen, dann reagieren.**

Nutze diese Checkliste bei überraschenden Inhalten. Selbst Expert*innen prüfen lieber doppelt, bevor sie Fehlinformation weiterverbreiten.

Details prüfen

- **KI-generierte Videos und Bilder können überzeugend aussehen, aber kleine Unstimmigkeiten verraten sie oft. Dieser erste Hinweis hilft dir, wahrzunehmen, dass etwas „nicht stimmt“, bevor du genau weißt, warum.**
- **Deine erste Detektivfähigkeit: langsamer werden und genauer hinsehen. Deepfakes übersehen oft Details, die im echten Leben normalerweise stimmen.**
- KI kann Gesichter, Stimmen und Gesten nachahmen, hat aber weiterhin Schwierigkeiten mit feinen Details. Solche Fehler können subtil sein: eine seltsame Spiegelung in einer Brille, ein zusätzlicher Finger, unpassende Schatten oder Ohrringe, die zwischen Bildern die Seite wechseln. Diese Hinweise zu bemerken ist der erste Schritt zu digitaler Intuition.

Quelle prüfen

- Suche nach der ursprünglichen Person, die den Inhalt hochgeladen hat, oder nach der Website, auf der er veröffentlicht wurde.
- Prüfe, ob vertrauenswürdige Nachrichtenmedien oder offizielle Konten dieselbe Geschichte berichten.
- Nutze eine umgekehrte Bildsuche, um zu sehen, ob das Bild anderswo in einem anderen Kontext auftaucht.
- Denke daran: Die erste Quelle, die du siehst, ist selten die zuverlässigste.

Quelle prüfen

Selbst das realistischste Bild oder Video kann irreführend sein, wenn es aus einer unzuverlässigen Quelle stammt. Dieser zweite Hinweis konzentriert sich darauf, wo ein Inhalt entstanden ist und wie du ihn überprüfen kannst.

Man vergisst leicht, dass jeder Post, jedes Bild und jedes Video eine Urheberin oder einen Urheber hat: jemand hat entschieden, es zu teilen. Ein Deepfake oder irreführendes Bild kann von einem Fake-Konto hochgeladen oder von jemandem weitergeteilt werden, der es selbst nie geprüft hat. Gute Detektiv*innen verfolgen die Spur zurück.

Gefühl prüfen

Du musst nicht sofort wissen, ob etwas echt ist. Aber du kannst immer hinterfragen, wie es auf dich wirkt.

- Manchmal ist dein eigenes Gefühl das beste Werkzeug. Dieser letzte Hinweis hilft dir zu erkennen, wann Emotionen manipuliert werden – eine zentrale Fähigkeit für digitale Resilienz.
- Wenn etwas zu perfekt, zu schockierend oder zu emotional wirkt, soll es dich vielleicht zum Reagieren bringen, bevor du nachdenkst.
- KI-generierte Inhalte spielen oft mit unseren Emotionen. Sie können dich wütend, inspiriert oder ängstlich machen.
- Emotionen bringen Menschen zum Klicken, Teilen und Glauben – genau deshalb werden sie gezielt angesprochen.

Gefühl prüfen

Notiere, was du darüber gelernt hast, emotionale Reaktionen vor dem Teilen zu verlangsamen.

Wenn etwas eine starke Reaktion auslöst, halte inne. Kritisches Denken bedeutet nicht Zynismus, sondern Neugier. Du musst nicht sofort wissen, ob etwas echt ist. Aber du kannst immer hinterfragen, wie es auf dich wirkt.

Denkst du darüber nach, Inhalte online zu teilen?

Prüfe zuerst diese 5 Schritte, um dich und andere online zu schützen!

Warum wir online kritisch denken müssen

Forschung zeigt:

Schon kurzer Kontakt mit falschen Inhalten kann Überzeugungen prägen
(Pennycook & Rand, 2019)

Emotionale Inhalte verbreiten sich schneller als faktische Korrekturen
(Vosoughi et al., 2018)

Widerlegung allein löscht die emotionale Wirkung nicht aus
(Nyhan & Reifler, 2010)

Faktencheck-Tools & Methoden

1. Umgekehrte Bildsuche – ein Bild rückwärts suchen, um die ursprüngliche Quelle, frühere Verwendungen und mögliche Kontextverschiebungen zu finden.
2. Metadaten prüfen – versteckte Dateidaten untersuchen, z. B. Erstellungsdatum, Gerät, Ort oder Autor*in, sofern verfügbar.
3. Mit offiziellen Kanälen vergleichen – Inhalte mit Aussagen von Institutionen, Organisationen, verifizierten Konten oder Original-Websites abgleichen.
4. Faktencheck-Portale – Websites nutzen, die Behauptungen, Bilder und Videos auf ihre Richtigkeit prüfen.

Faktencheck-Tools & Methoden

5. Primärquellen prüfen – Informationen bis zur ursprünglichen Quelle zurückverfolgen, statt sich auf Reposts, Zitate oder Zusammenfassungen zu verlassen.
6. Veröffentlichungsdatum und -ort prüfen – kontrollieren, wann und wo Material veröffentlicht wurde und ob es zum beschriebenen Ereignis passt.
7. Kontextanalyse – den vollständigen Kommunikationskontext prüfen, um Zitate, Bilder oder Videos zu erkennen, die aus ihrem ursprünglichen Zusammenhang gelöst wurden.
8. Visuelle Geolokalisierung – Elemente in Foto oder Video mit Karten, Gebäuden, Schildern, Landschaften oder Wetterbedingungen vergleichen, um den Aufnahmeort zu bestimmen.

05

Verantwortung von Content-Ersteller*innen und Folgen von Desinformation

(15 Min.)

Verantwortung von Content-Ersteller*innen

1. Veröffentliche kein ungeprüftes Material – prüfe vor dem Teilen, ob die Information stimmt und von mehreren zuverlässigen Quellen bestätigt wurde.
2. Teile keine Inhalte, deren Echtheit du nicht kennst – wenn du nicht sicher bist, ob das Material echt ist, verstärke es nicht weiter.
3. Kennzeichne synthetische Inhalte – mache klar kenntlich, wenn ein Bild, Audio, Text oder Video mit KI erstellt oder verändert wurde.
4. Korrigiere Fehler – wenn du einen Fehler in veröffentlichten Inhalten bemerkst, korrigiere ihn so schnell und sichtbar wie möglich.
5. Entferne oder korrigiere irreführendes Material – wenn Inhalte falsche Tatsachen nahelegen, sollten sie entfernt oder klargestellt werden.

Verantwortung von Content-Ersteller*innen

6. **Nutze Bild oder Stimme einer Person nicht ohne Einwilligung – Gesicht, Stimme oder Sprechweise einer anderen Person ohne Erlaubnis zu verwenden, kann Privatsphäre und Persönlichkeitsrechte verletzen.**
7. **Gib Quellen und Kontext an – erkläre, woher das Material stammt, wann es erstellt wurde und warum es veröffentlicht wird.**
8. **Manipuliere das Publikum nicht absichtlich – Ersteller*innen sollten Handlungen vermeiden, die die Öffentlichkeit bewusst in die Irre führen, um Reichweite, Gewinn oder emotionale Wirkung zu erzielen.**

Auswirkungen irreführender KI

Gefälschte oder irreführende KI-generierte Inhalte enden nicht bei einer einzelnen Person.

Wenn ein Fake-Bild, Video oder eine Geschichte geteilt wird, verbreitet sich der Inhalt schnell über Plattformen hinweg, erreicht Menschen, die dem Absender vertrauen, und kann weit über den ursprünglichen Post hinaus Schaden verursachen.

„Forschende beschreiben dies als Welleneffekt: Eine einzelne Fehlinformation kann Vertrauen, Verhalten und Entscheidungen in großem Maßstab beeinflussen.“

(Wardle & Derakhshan, 2017).

Arten von Schaden

KI-generierte Fälschungen können verursachen:

- **Persönlichen Schaden: Rufschädigung oder mentale Belastung**
- **Sozialen Schaden: Angst, Konflikte, Diskriminierung**
- **Demokratischen Schaden: Wähler*innen täuschen, Vertrauen schwächen**
- **Sicherheitsrisiken: Betrug oder falsche medizinische Informationen**

Arten von Schaden

Auch wenn Teilen harmlos wirkt, ist die Wirkung oft nicht harmlos. Schaut euch „Bad science: AI used to target kids with disinformation on YouTube – BBC World Service“ an.

06

Übungen und Beispiele

(25 Min.)

Übung: Rollenspiel – Das Deepfake-Dilemma

Szenario 01

Ein Deepfake einer Mitschülerin oder eines Mitschülers wird in sozialen Medien „als Witz“ geteilt – was sollten Freund*innen tun?

Szenario 02

Ein beliebter Online-Influencer postet ein Video mit seiner Partnerin oder seinem Partner. Später stellt sich heraus: Diese Person existiert nicht und wurde vollständig mit KI erstellt, um mehr Aufrufe zu erzielen. Fans fühlen sich betrogen.

Übung: Rollenspiel – Das Deepfake-Dilemma

Szenario 03

Nach einer chaotischen Trennung teilt jemand online ein Video, das scheinbar zeigt, wie die Ex-Partnerin oder der Ex-Partner grausame Dinge sagt oder einen Betrug zugibt. Der Clip verbreitet sich schnell – ist aber tatsächlich ein KI-generierter Deepfake.

Szenario 04

Du entdeckst, dass jemand deine Stimme in einem Prank-Video mit KI imitiert hat – wie gehst du damit um?

„Papst in Daunenjacke“

Fake-Fotos von Papst Franziskus in einer Daunenjacke gehen viral und zeigen Macht und Risiko von KI

CBS News

„Papst in Daunenjacke“

- Was passiert ist: Ein KI-Bild von Papst Franziskus in einer weißen Daunenjacke ging viral; viele hielten es für echt.
- Was täuscht: Fotorealismus, Anknüpfung an Modetrends.
- Absicht: Spielerisch.
- Schaden: Vertrauen wird geschwächt; Nicht-Prüfen wird normalisiert.

Welche kleinen Hinweise könntest du in diesem Bild prüfen, bevor du es teilst?

TikTok-Filter „Bold Glamour“

TikToks „Bold Glamour“-Beauty-Filter beunruhigt KI-Expert*innen mit Blick auf Realität

The Washington Post

TikTok-Filter „Bold Glamour“

- **Was passiert ist: Ein hyperrealistischer Beauty-Filter nutzt maschinelles Lernen, um Gesichter live zu verändern.**
- **Was täuscht: Nahtlose KI-Gesichtsbearbeitung wirkt „natürlich“.**
- **Absicht: Unterhaltung/Engagement.**
- **Schaden: Unrealistische Standards und Druck auf das Körperbild.**

Wie könnte dieser Filter beeinflussen, wie Menschen sich selbst und andere sehen?

„Deepfake-CFO im Zoom-Call“

Mitarbeiter eines Unternehmens in Hongkong zahlt 20 Mio. Pfund nach Deepfake-Videoanruf-Betrug aus

Hongkong | The Guardian

„Deepfake-CFO im Zoom-Call“

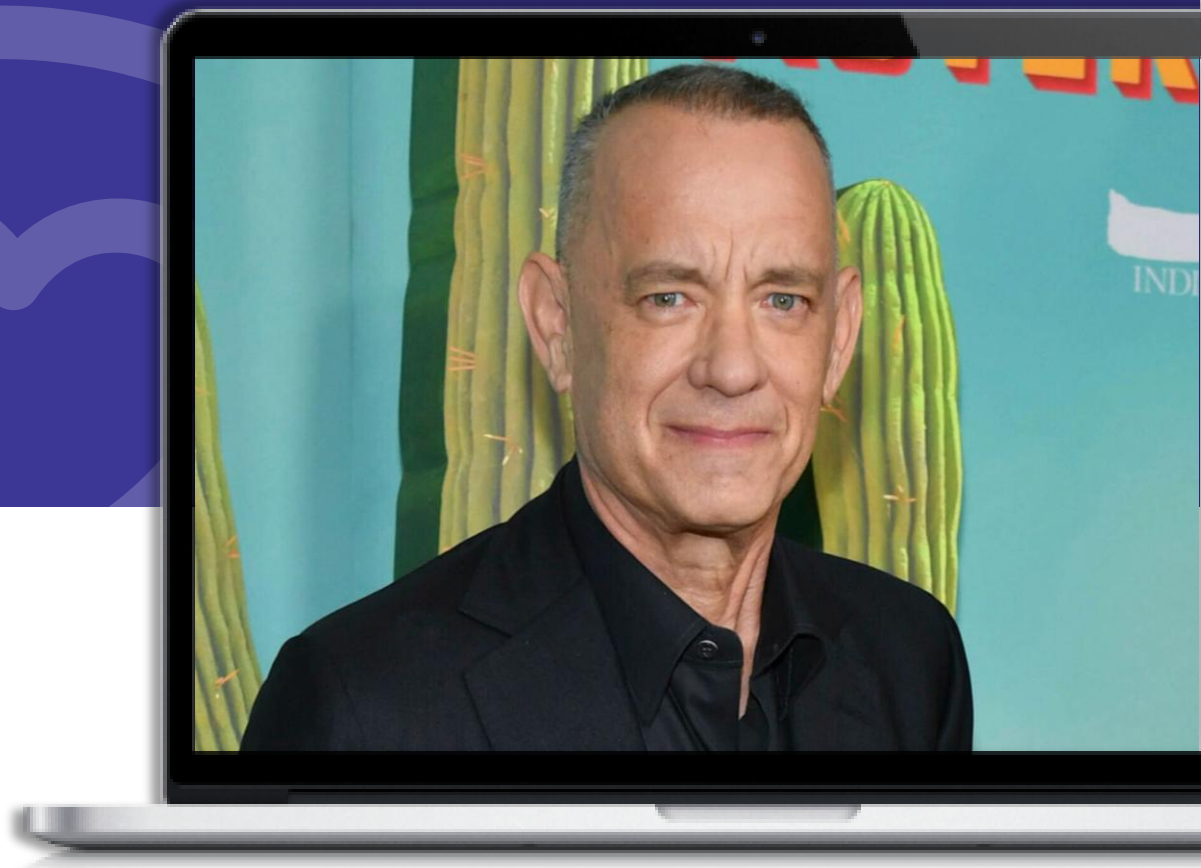
- **Was passiert ist: Ein Finanzmitarbeiter in Hongkong wurde durch einen Deepfake-Videoanruf mit mehreren „Kolleg*innen“ getäuscht und überwies rund 25 Mio. US-Dollar.**
- **Was täuscht: Überzeugende Video- und Stimmklone bekannter Mitarbeitender.**
- **Absicht: Betrug.**
- **Schaden: Hoher finanzieller Verlust, Reputations- und Rechtsrisiko.**

Nenne zwei externe Prüfwege, die du vor jeder Überweisung verlangen würdest.

„Diese Werbung bin nicht ich“

Tom Hanks warnt Fans vor einer Deepfake-Werbung, die sein Gesicht nutzt

Ents & Arts News | Sky News



„Diese Werbung bin nicht ich“

- **Was passiert ist: Ein KI-Video nutzte Tom Hanks’ Erscheinungsbild, um ein Produkt zu verkaufen; er warnte seine Fans.**
- **Was täuscht: Vertrautes Gesicht und vertraute Stimme, Werbeformat.**
- **Absicht: Kommerzielle Täuschung.**
- **Schaden: Verbraucherbetrug, Missbrauch von Bildrechten.**

Wie würdest du eine Promi-Empfehlung überprüfen, bevor du sie glaubst?

„Heute nicht wählen“

Politikberater hinter KI-generierten Biden-Robocalls drohen 6 Mio. Dollar Geldstrafe und Strafanzeigen

AP News

„Heute nicht wählen“

- **Was passiert ist: Eine KI-Stimme, die Präsident Biden imitierte, sagte Wähler*innen, sie sollten nicht an der Vorwahl teilnehmen; Geldstrafen und Anklagen folgten.**
- **Was täuscht: Vertraute politische Stimme, kurz vor der Wahl.**
- **Absicht: Wählerunterdrückung.**
- **Schaden: Beeinträchtigung demokratischer Teilhabe.**

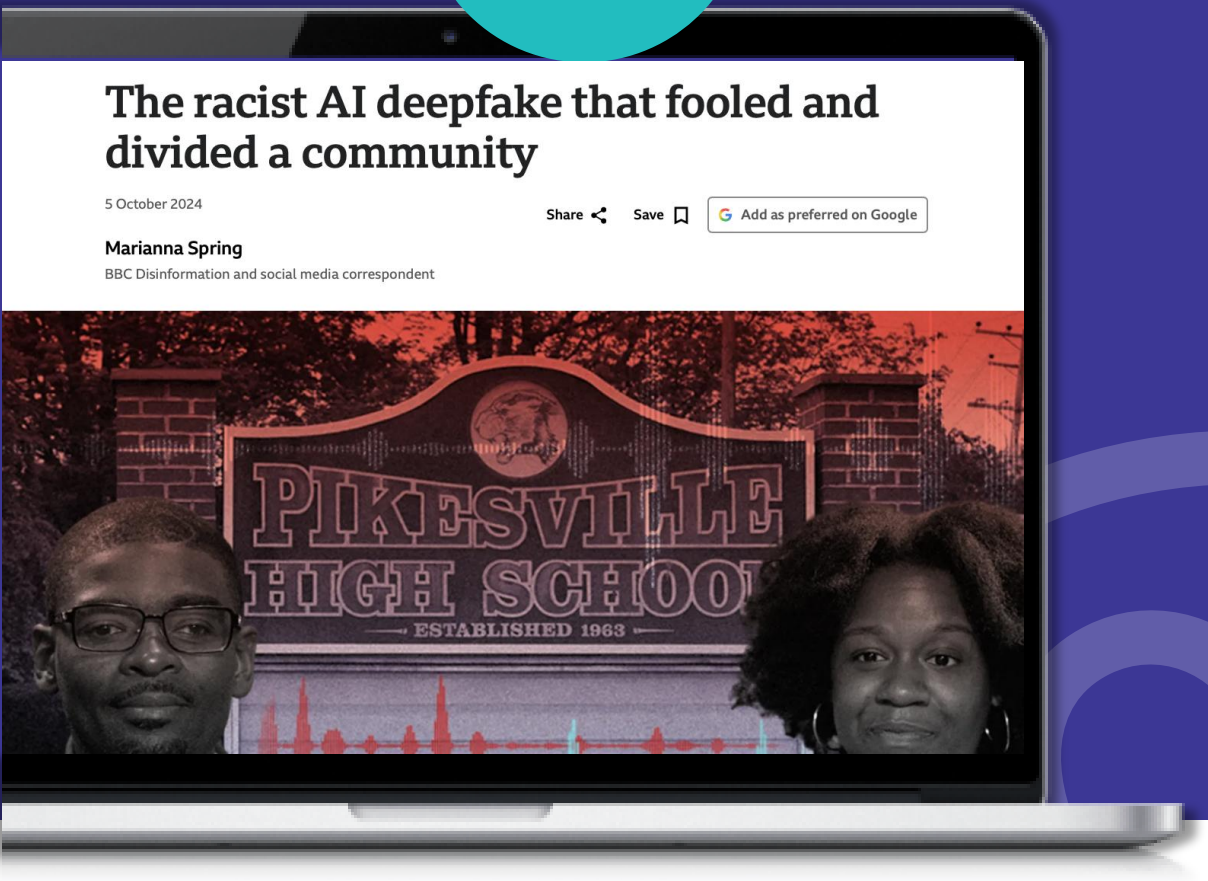
Was sollte ein*e Wähler*in tun, wenn er oder sie einen solchen Anruf erhält?

Übung: Rollenspiel – Das Deepfake-Dilemma

1. Teilt euch in 3–4 Kleingruppen auf.
2. Jede Gruppe spielt ein Szenario durch.
3. Die Darstellung dauert höchstens
4. 2–3 Minuten pro Gruppe.
5. Anschließend folgt eine Gruppenreflexion.

Fallstudie: „Das Fake-Audio, das eine Gemeinschaft spaltete“

CLICK TO VIEW

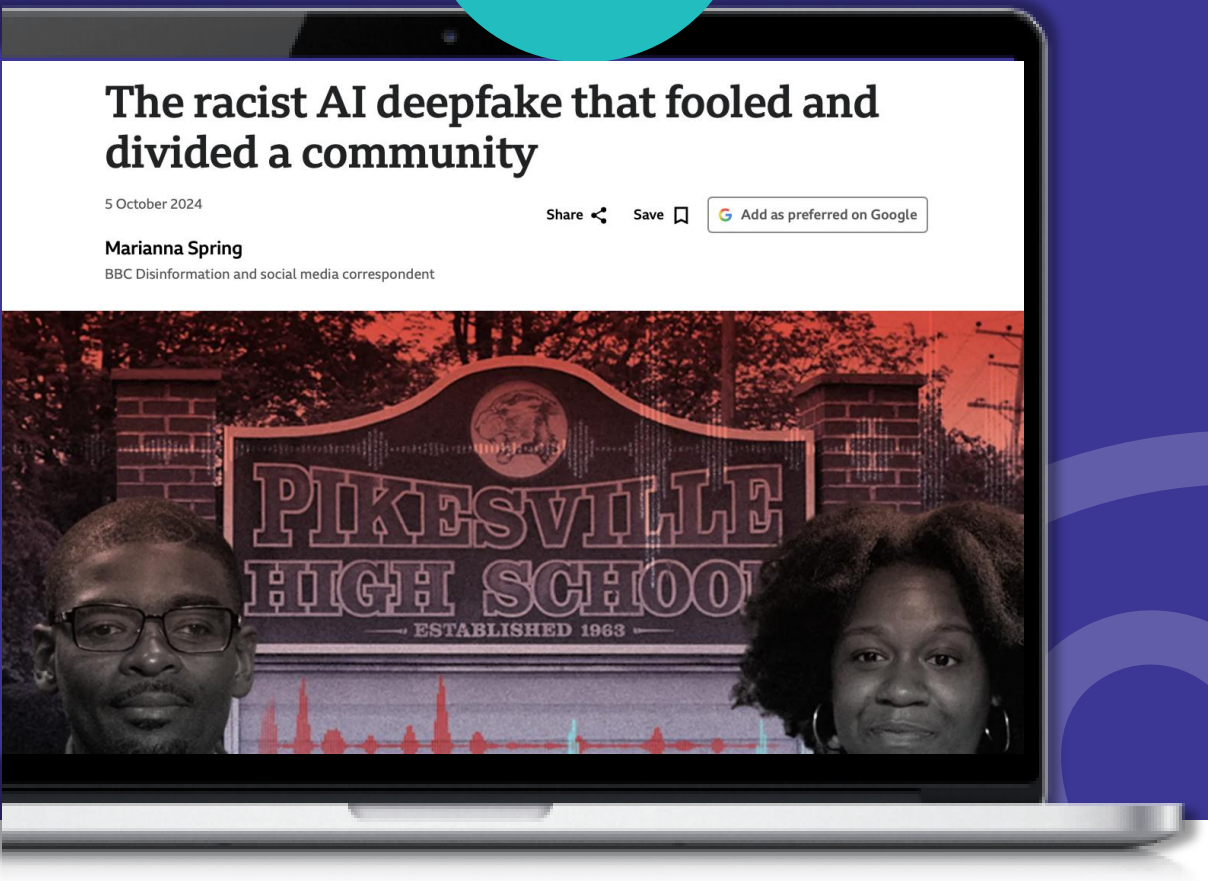


In einer Stadt in der Nähe von Baltimore (USA) tauchte online ein Audioclip auf, der so klang, als würde ein Schulleiter rassistische und antisemitische Kommentare machen.

- Der Clip verbreitete sich schnell in sozialen Medien.
- Viele Menschen hielten ihn für echt.
- Eltern waren wütend.
- Die Schule erhielt Drohungen.

Fallstudie: „Das Fake-Audio, das eine Gemeinschaft spaltete“

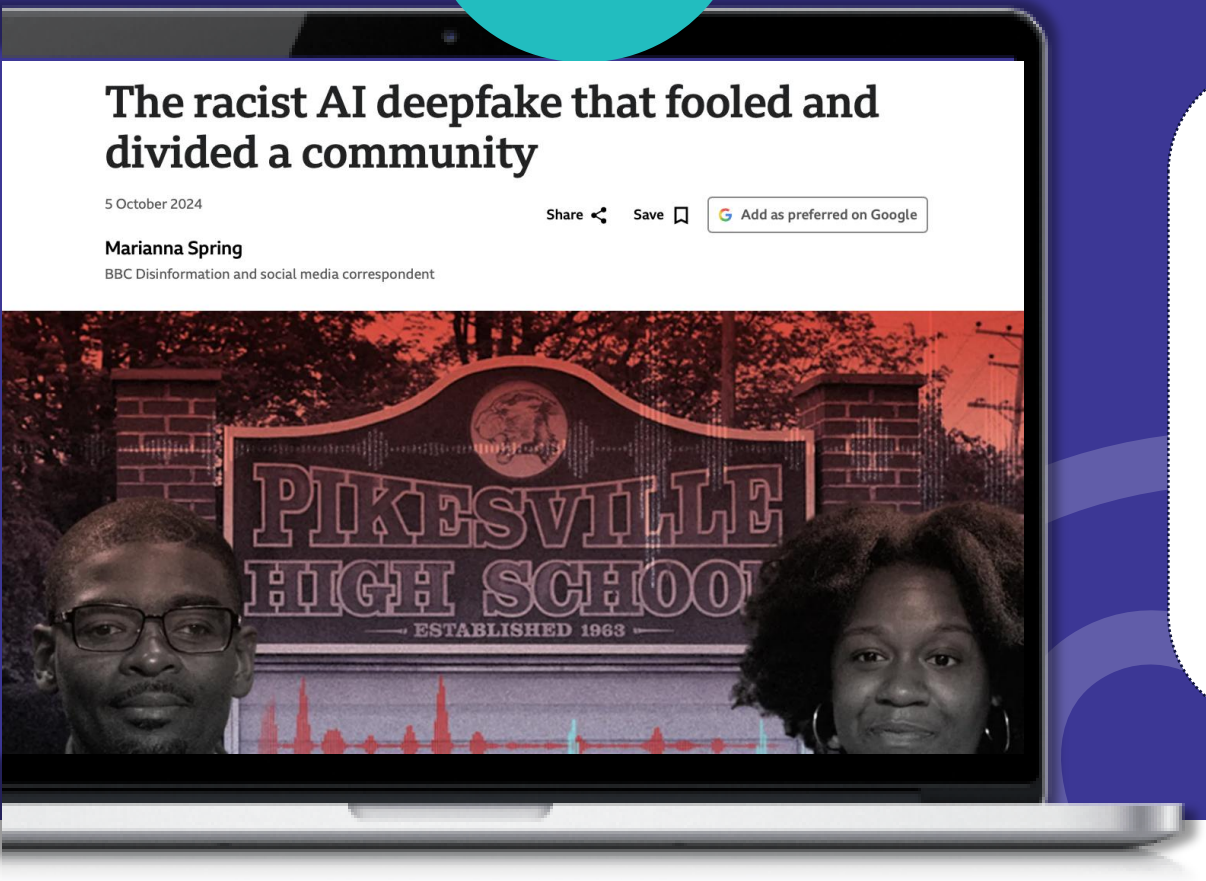
CLICK TO VIEW



Der Schulleiter wurde während der Ermittlungen beurlaubt.
Später bestätigte die Polizei, dass das Audio KI-generiert und gefälscht war. Doch selbst nachdem dies bewiesen war, blieb der Schaden bestehen: Das Vertrauen in die Schule war beschädigt, Menschen fühlten sich weiterhin verletzt und wütend, und manche glaubten weiter, der Clip zeige, „wie Menschen wirklich denken“.

Fallstudie: „Das Fake-Audio, das eine Gemeinschaft spaltete“

CLICK TO VIEW



Der Schulleiter wurde während der Ermittlungen beurlaubt. Später bestätigte die Polizei, dass das Audio KI-generiert und gefälscht war. Doch selbst nachdem dies bewiesen war, blieb der Schaden bestehen: Das Vertrauen in die Schule war beschädigt, Menschen fühlten sich weiterhin verletzt und wütend, und manche glaubten weiter, der Clip zeige, „wie Menschen wirklich denken“.

Fallstudie: „Das Fake-Audio, das eine Gemeinschaft spaltete“

Warum glaubten Menschen daran?
Der Clip wirkte glaubwürdig, weil:

- er authentisch klang (Stimmklon),
- er Insiderdetails enthielt (Schulsprache, Namen),
- er zu bestehenden Ängsten passte,
- Audio weniger visuelle Hinweise bietet,
- Plattformen ihn vor der Prüfung verstärkten.

Authentizitäts-Bias: Wenn etwas echt wirkt, akzeptieren wir es leichter als echt. KI-Fakes nutzen Vertrauen, Emotionen und frühere Erfahrungen aus. Kritisches Denken heißt deshalb auch: Gefühle und Vorannahmen prüfen.

Fallstudie: „Das Fake-Audio, das eine Gemeinschaft spaltete“

Was hätte den Schaden verlangsamen können?
Diskutiert gemeinsam:

- An welchem Punkt hätte sich die Verbreitung verlangsamen lassen?
- Vor dem Teilen? Vor emotionalen Reaktionen? Bevor Plattformen den Inhalt verstärkten?
- Welche Prüffragen sind hier am wichtigsten?
  - Wahr? (War die Quelle verifiziert?)
  - Absicht? (Wer hat es gesendet und warum?)
  - Achtsam? (Wer könnte geschädigt werden?)
- Warum war „es fühlt sich wahr an“ stärker als „es ist wahr“?

Evidenz-Schnappschuss: dein „Detektiv-Protokoll“

- Du hast 5–8 Minuten für diese Aktivität – allein oder zu zweit.
- Wähle einen Inhalt aus (Bild, Video, Audio oder Chatbot-Antwort).
- Erstelle eine kurze „Detektiv-Notiz“ mit der Vorlage unten.
- Vorlage Detektiv-Notiz (ausfüllen)
- Was ist der Inhalt? (1 Satz)
- Wo habe ich ihn gefunden? (Plattform/Konto/Link, wenn möglich)
- Meine erste Reaktion: (Was habe ich gefühlt?)

Evidenz-Schnappschuss: dein „Detektiv-Protokoll“

Meine Checks

- ❑ Detail-Check (Was wirkte merkwürdig?)
- ❑ Quellen-Check (Wer hat gepostet? Gibt es vertrauenswürdige Quellen?)
- ❑ Emotions-Check (Sollte ich zu einer Reaktion gedrängt werden?)

Entscheidung

- ❑ Ich vertraue dem Inhalt genug, um ihn zu nutzen/teilen
- ❑ Ich bin unsicher (ich teile noch nicht)
- ❑ Ich halte ihn für irreführend/gefälscht

Begründung (1 Satz):

Warum wichtig? Deine Begründung macht dein Denken sichtbar – ein Kern von kritischem Denken und Metakognition.

AI Skills,
Powered
by Values

values

Folge unserer Reise

www.valuesai.eu

Co-funded by
the European Union

Mitfinanziert von der Europäischen Union. Die geäußerten Ansichten und Meinungen sind jedoch ausschließlich die der Autor*innen und spiegeln nicht unbedingt die Ansichten der Europäischen Union oder der Stiftung für die Entwicklung des Bildungssystems wider. Weder die Europäische Union noch die fördernde Einrichtung können dafür verantwortlich gemacht werden.