EFCSN

Wie man KI-generierte oder digital veränderte Inhalte erkennt

Leicht zugängliche und nutzerfreundliche KI-Modelle können beim Lernen und beim Erstellen von Inhalten helfen. Sie können aber auch die **Risiken vergrößern, die Desinformation und Fehlinformationen** für offene Gesellschaften und den demokratischen Austausch bedeuten. Es ist wichtig, unsere gemeinsamen Räume für den Austausch von Informationen vor einer Überflutung mit KI-generierten und digital veränderten Falschinformationen zu schützen.

Teil der Lösung sind neue Technologien, etwa Initiativen zur Kennzeichnung der Herkunft von Inhalten oder Software zum Erkennen solcher Inhalte. **Aber technologische Lösungen sind längst nicht perfekt. Wir brauchen die Arbeit unabhängiger Faktenchecker, um für die Gesellschaft eine gemeinsame Basis verifizierter Fakten zu schaffen.**

Hier ist ein Überblick darüber, wie unabhängige Faktencheck-Profis KI-generierte Falschinformation entlarven, und was du von ihnen lernen kannst.

KI-generierte Inhalte nehmen zu

Derzeit machen KI-generierte Falschinformationen einen kleinen Teil aller von professionellen, unabhängigen Faktencheckerinnen und Faktencheckern untersuchten Behauptungen aus. Digital veränderte Inhalte sind präsenter.

Aber in einer internen Umfrage unter den Mitgliedern des EFCSN waren sich die meisten einig, dass KI-generierte und digital veränderte Inhalte wichtiger werden. Aktuelle Beispiele rund um die Europawahlen bestätigen das.

KURZ ERKLÄRT: *Digital verändert* meint Inhalte, deren ursprüngliche Aussage verändert wurde, auch mit Hilfe von KI-Tools. Bearbeitungen, die die Qualität verbessern oder etwas verdeutlichen sollen, gehören nicht dazu.

KI-generiert meint Inhalte, die von einem System künstlicher Intelligenz geschaffen wurden.

Diese Informationen stammen von **AI@EUElections,** gefördert von Meta & verantwortet vom **European Fact-Checking Standards Network** mit seinen Mitgliedern und externen Fachleuten.

Technologie entwickelt sich schnell, aber wir können uns nicht nur darauf stützen.

KI-Experten and Faktencheck-Profis sind sich einig: **Tools zur KI-Erkennung allein reichen nicht, um KI-generierte oder digital veränderte Inhalte zu entlarven.**

Fachleute empfehlen, sich zunächst mit Tools zum Erstellen und Erkennen von KI-Inhalten vertraut zu machen. Wenn Faktenchecker verstehen, wie die Modelle lernen, können sie ein Verständnis für Stärken und Schwächen von Tools entwickeln und ihre Erfolgschancen einschätzen. **Dennoch können Tools ein guter Ausgangspunkt sein.**

Initiativen zur **Kennzeichnung der Herkunft von Inhalten**, etwa C2PA, können helfen, Quelle und Bearbeitungs-Historie nachvollziehbar zu machen. Wasserzeichen und Nachweise sind aber nicht unveränderlich.

Wie KI-Desinformation auf Menschen wirkt

"Jedes Mal, wenn Sie aus dem Bauch heraus reagieren, umgehen Sie das Nachdenken."
Christine Dugoin*

PSYCHOLOGIE: Kampagnen zur Beeinflussung sollen oft gezielt psychologische Verzerrungen nutzen.

Diese bei sich selbst und dem Gegenüber zu verstehen, kann gegen Desinformation helfen.

ABSICHTEN: Warum könnte jemand mit schlechten Absichten KI nutzen, um Desinformation zu erfinden oder zu verbreiten? Was soll das in der echten Welt bewirken?

- ➔ Ein anderes Land oder eine andere Community erreichen?
- ➔ Nicht auffliegen oder Faktenchecker mit Variationen der selben Behauptung überschütten?
- ➔ Meinungen beeinflussen, indem ein Netzwerk falscher Accounts glaubwürdig wirkt?

** Christine Dugoin erforscht an der Sorbonne die Beeinflussung mit Hilfe von Informationen.*

Widerlegen erfordert einen vielschichtigen Ansatz und ein detailliertes Verständnis

Wenn Tools zur KI-Erkennung nicht funktionieren, was dann? Der Kontext einer Behauptung ist so wichtig wie ihr Inhalt. Faktencheck-Profis haben die nötigen Recherche-Kenntnisse. Hier sind ein paar Tipps.

> *"Erkennungs-Tools werden nie zu 100% funktionieren - ich denke, das passiert nicht."* - Henk van Ess**

BERÜCKSICHTIGE DIE QUELLE: Kannst du ihre Identität bestätigen? Wozu äußert sie sich, was teilt sie? Wer reagiert auf ihre Inhalte? Welche Wirkung könnte dieser Inhalt auf Lesende haben?

PRÜFE DIE GLAUBWÜRDIGKEIT: Prüfe die Informationen unabhängig anhand glaubwürdiger Quellen, etwa Fachleute mit Erfahrung auf diesem Gebiet. Ergibt das Dargestellte deinem Wissen nach Sinn?

Nutze **MEDIENFORENSIK**-Techniken, um traditionelle investigative und dokumentarische Recherchen zu ergänzen. Dazu gehören: Daten-Scraping, Ortsbestimmung, biometrische Gesichtserkennung oder Musteranalyse.

LERNE & SEI FLEXIBEL: KI-generierte Falschinformationen sind ständig im Wandel. Passe deine Methoden entsprechend an.

LASS ANDERE AN DEINER ARBEIT TEILHABEN

Expertinnen und Experten empfehlen, zu der widerlegten Behauptung eine transparente Analyse zu stellen und auf Quellen zu verlinken. Das kann Leserinnen und Lesern helfen, eine Recherche nachzuvollziehen und Nuancen zu verstehen. Manchmal ist die Recherche wichtiger als die Frage, ob der Inhalt von einer KI geschrieben wurde.

*** Henk van Ess ist Experte für OSINT- und Faktencheck-Techniken.*

Kurzanleitung: Dos, Don'ts und Tipps

Dies sind Anzeichen dafür, dass ein Inhalt KI-generiert oder digital verändert sein könnte. Zusammen mit den anderen Tipps (Kontext, Ermittlungstechniken und Erkennungstools) können sie helfen, die Fakten hinter dem zu verstehen, was du siehst.

Text

- ➔ Oft (nicht immer) ist die **Grammatik** besser als in Texten von Menschen.
- ➔ Häufig ungewöhnlich **formelle oder strukturierte Sprache**, vor allem für Social Media.
- ➔ Besonders viele **Adverbien oder Adjektive**.
- ➔ Fehlen von menschlichen Gefühlen, Humor, Sarkasmus und Redewendungen.
- ➔ Es können **Details fehlen** (Namen, Daten, Orte) oder originelle Ideen.
- ➔ Vor allem: Stimmen die Behauptungen im Text?

Video

- ➔ Nutze keinen Detektor für KI-Bilder zum Prüfen von Standbildern aus Videos.
- ➔ Achte auf **Gesichtsausdrücke und Bewegungen** wie Blinzeln, und ob Mund und Ton zusammenpassen.
- ➔ Möglicherweise **abrupte Übergänge oder Schnitte**.

Audio

- ➔ **Vergleiche verdächtige Tonspuren** mit einem echten Beispiel mit Hilfe von Tools, die Unterschiede erkennen im Sprach- und Atemmuster, der Betonung etc.
- ➔ Wenn du Erkennungs-Tools nutzt, vermeide Tonspuren in geringer Qualität mit Stör- oder Hintergrundgeräuschen.
- ➔ Möglicherweise **unnatürliche, mechanische Sprachmuster** oder fehlende Atempausen.

Bilder

- ➔ Suche nach **unnatürlichen Details**: perfekte Haut, unscharfer Hintergrund, künstliche Schönheit oder Belichtung, Kurioses wie zusätzliche Finger.
- ➔ **Achte auf Wasserzeichen** von gängigen Bild-Generatoren.
- ➔ Beachte Details: Sind sie logisch? Sind sie passend?
- ➔ Wenn du Erkennungs-Tools nutzt, **wähle besser eine hochauflösende oder frühe Version** als eine schon vielfach geteilte.