68. JAHRGANG NR. 3 / JUNI 2024

€ 10,–

merz

MEDIEN + ERZIEHUNG

ZEITSCHRIFT FÜR MEDIENPÄDAGOGIK

MEDIENPÄDAGOGIK UND KI

WEITERE THEMEN:

METAVERSUM UND KINDERRECHTE

FRAGWÜRDIGE GESTALTEN – FRAGWÜRDIGES GESTALTEN

AUS DEN FORSCHUNGSWERKSTÄTTEN VON ‚DAS BEWEGT UNS'

kopaed

Was steckt hinter generativer KI und wie kann man sie erkennen? Wie kann die verantwortungsvolle Nutzung ihrer unendlichen kreativen Möglichkeiten gelingen? Und welche Tools können dabei helfen, KI generierte Werke von menschlichen zu unterscheiden? Im folgenden Text werden konkrete Hilfestellungen gegeben, deren Nutzung dazu beitragen kann, die Verbreitung von Fehlinformationen einzudämmen und faktenbasierte Debatten zu erleichtern.

KI-GENERIERTE BILDER, TEXTE UND VIDEOS ERKENNEN

Lea Uhlenbrock

Genau wie einst das Internet und danach Smartphones erobert KI nun nach und nach unseren Alltag und ist nicht mehr wegzudenken. Vom automatisierten Social-Media-Algorithmus, der passgenaue Vorschläge in den *Instagram*-Feed spült, bis zur Spracherkennung auf dem Smartphone übernimmt sie immer mehr Aufgaben. Eine besonders spannende Entwicklung ist dabei die von sogenannter generativer KI, also KI, die digitale Werke erzeugen kann. Das sind beispielsweise Texte von *ChatGPT*, Bilder von *Midjourney* oder Videos von *Sora*.

Generative KI funktioniert wie andere KI-Formen: Aus einer großen Menge Daten werden Muster gelernt und abstrakt abgespeichert. Das Besondere bei dieser KI-Form ist, dass die gelernten Muster in neue Inhalte übersetzt werden können. Es werden also Eigenschaften von typischen Texten und Bildern gelernt und mit einem gewissen Anteil an Zufall als ‚neuer' Inhalt ausgegeben.

Bilder, die von frühen Versionen solcher KI generiert wurden, wie *DALL-E mini* im Sommer 2022, waren noch leicht als unecht zu entlarven, da sie grobe visuelle Fehler beinhalteten. Durch die rapide Evolution von Bildgeneratoren hinweg wird es allerdings kontinuierlich schwieriger, synthetische Bilder zu erkennen. Während bei *Midjourney* 4 und 5 noch oft eine seltsame Anzahl an Fingern an menschlichen Händen generiert wurde, sind Hände auf Bildern von *Midjourney* 6 kaum mehr von echten zu unterscheiden. Auch andere visuelle Fehler tauchen deutlich seltener auf, weil die Bildgeneratoren besser werden und lernen, solche ‚störenden' Bildmakel zu vermeiden.

Die Möglichkeit, beliebige Bildinhalte mit der einfachen Eingabe eines Beschreibungstextes, eines sogenannten Prompts, zu generieren, schafft vielfältige kreative Möglichkeiten. Storyboards für Filme können mit wenigen Klicks erstellt werden, genau so wie Abbildungen von Fantasiewelten und Charaktere für ein Pen and Paper Rollenspiel oder Bilder für einen Veranstaltungsflyer. Doch die unbegrenzten Möglichkeiten von generativer KI beinhalten auch Gefahrenpotenzial. Parteien mit extremen Ansichten nutzen sie bereits auf *Instagram*, um Propaganda zu verbreiten. Beobachtet man im öffentlichen Generierungs-Kanal von *Midjourney*, welche Inhalte erstellt werden, kann man viel Politisches entdecken, mit gezielten Botschaften und in teilweise sehr überzeugender Fotoqualität. Passend dazu generierte Texte machen die Verbreitung von Fake News oder Hetze einfacher denn je.

Gerade bei Bildern und Videos – Medien, denen wir Menschen besondere Glaubhaftigkeit zuordnen, ist es wichtig, solche Inhalte zuverlässig und automatisiert entlarven zu können. Die

Wissenschaft arbeitet deshalb intensiv daran, solche generierten Inhalte automatisch zu erkennen. Bilder lassen sich beispielsweise schon jetzt oft anhand ihrer Rauschmuster erkennen. Echte Kameras hinterlassen wissenschaftlich erkennbare Rauschmuster in Fotos, die in KI-generierten Bildern so nicht vorkommen. Auch generierte Texte lassen sich oft anhand bestimmter Muster erkennen und entlarven. Es gibt bereits einige automatisierte Tools, die das Erkennen von generierten Texten und Bildern ermöglichen.

Für Texte gibt es beispielsweise:

- https://GPTZero.com
- https://copyleaks.com
- https://DetectGPT.com

Für Bilder gibt es unter anderem:

- https://AIorNot.com
- https://IsItAI.com
- https://illuminarty.ai

Solche Werkzeuge arbeiten jedoch nicht immer zuverlässig oder fehlerfrei. Es lohnt sich daher, Bilder, Texte und Videos genau unter die Lupe zu nehmen und nach Anzeichen zu suchen, ob sie eventuell generiert sind. Bei Bildern sind beispielsweise in den Details oft noch Fehler in der Bildlogik versteckt, die so in der Realität nicht vorkommen. Der genaue Blick auf reale Gegebenheiten wie Licht, Schatten, Spiegelungen und Geometrie ist wichtig, um zu erkennen, wann etwas falsch oder unlogisch dargestellt wird. Hilfreich ist es hier vor allem, einzelne Objekte nachzuverfolgen und genau zu betrachten. Wenn auf einem Bild beispielsweise eine Person gezeigt wird, die eine Kette trägt, kann man der Kette folgen und wird bei synthetischen Bildern häufig feststellen, dass sie irgendwann in die Kleidung übergeht oder sich mit anderen Objekten geradezu vermischt. Bei Videos gilt Ähnliches, auch hier ist es wichtig, auf Unstimmigkeiten und Details zu achten. Bis jetzt besteht ein Großteil an KI-Videos noch hauptsächlich aus realem Bildmaterial, und nur das Gesicht der abgebildeten Person wurde verändert. Dabei kann man oft am Rand des Gesichts Phänomene wie Verformungen oder Verzerrungen beobachten, oder fehlende Veränderungen des Schattenwurfs, wenn sich die Person bewegt. Neuere Videos können auch vollständig KI-generiert sein und weisen dann ähnliche Fehler wie Bilder auf: Objekte vermischen sich innerhalb weniger

> BEOBACHTET MAN IM ÖFFENTLICHEN GENERIERUNGSKANAL VON MIDJOURNEY, WELCHE INHALTE ERSTELLT WERDEN, KANN MAN VIEL POLITISCHES ENTDECKEN, MIT GEZIELTEN BOTSCHAFTEN UND IN TEILWEISE SEHR ÜBERZEUGENDER FOTOQUALITÄT

Prompt: An artificial intelligence painting an image onto a canvas, photograph
// Midjourney

Video-Frames mit anderen Objekten, ändern ihre Form oder verschwinden sogar.
Bei Texten ist es nicht ganz so einfach, synthetische Inhalte anhand von Details zu erkennen. Hier zählt eher der Gesamteindruck. Generierte Texte basieren darauf, welche Wörter in der realen Welt mit der größten Wahrscheinlichkeit hintereinander auftauchen. Sie sind, wie eingangs erwähnt, nur eine Imitation von Sprachmustern und wirken deshalb auch genau so. Texte von *ChatGPT* beispielsweise wirken in der Regel unpersönlich, generisch und nicht wirklich konkret. Sie enthalten häufig falsche Angaben, die als Fakten dargestellt werden, oder erklären feststehende Konzepte nicht richtig. Gerade *ChatGPT* neigt außerdem dazu, Texte sehr positiv und bestätigend zu formulieren, was sie beinahe wirken lässt, als seien sie von einem Marketingteam verfasst. Der wichtigste Hinweis auf Urheber*innenschaft einer KI bei Texten ist der Kontext. Wenn der Text nicht zu Situation, Fragestellung oder Umständen passt oder der Stil völlig anders ist, als man es von der angegebenen Quelle kennt, können das Anhaltspunkte dafür sein, dass es sich um synthetischen Text einer KI handelt.
Im Folgenden finden sich drei Checklisten für Bilder, Texte und Videos, die dabei helfen sollen, generierte Inhalte zu erkennen. Wenn Sie alle Fragen mit Ja beantworten können, ist die Wahrscheinlichkeit, dass es sich um ein echtes Bild oder Video oder einen authentischen Text handelt, relativ hoch. Natürlich bieten die Checklisten keine Garantie, denn gut generierte Inhalte können sehr realistisch wirken und auch aufmerksame Beobachter*innen austricksen. Dennoch lohnt es sich, genauer hinzusehen und die Augen offenzuhalten.
Durch bewusstes Hinsehen und Betrachtung aller Inhalte in ihrem Kontext und mit einer gesunden Skepsis, sowie weiterer Forschung daran, generierte Inhalte aufzudecken, können wir dazu beitragen, die Verbreitung von Fehlinformationen einzudämmen und eine informierte Gesellschaft und faktenbasierte Debatte zu fördern.

BILDER

- Raum und Formen: Kann ich einzelne Objekte wie Ketten, Stäbe, Schnüre, Kleidungsstücke, Arme nachverfolgen, ohne dass sie die Farbe, Textur oder Form ändern oder ineinander übergehen?

- Einheitlichkeit: Sind Augenfarbe, Ohrringe, Kleidungsmerkmale für rechts und links gleich? Stimmen Eigenschaften von größeren Objekten wie Autos, Häusern, Zimmern symmetrisch überein?
- Anatomie: Haben alle abgebildeten Menschen oder Tiere die erwartete Anzahl Hände, Beine, Arme, Finger?
- Licht und Schatten: Verlaufen die Schatten alle von der Lichtquelle weg und an Hindernissen entlang logisch? Spiegeln sich Lichter an glatten Flächen? Bilden Spiegelungen das Original korrekt ab und stimmt die Perspektive?
- Individualität: Enthält das Bild individualisierende Elemente, Makel, Besonderheiten?
- Texturen: Sind Muster von Böden, Wänden, Maserungen konsistent? Sind glatte Oberflächen auch wirklich glatt, oder zeigen sie feine Muster, die einer Elefantenhaut ähneln?
- Ästhetik: Kann die Szene, wie sie stattgefunden haben soll, so ästhetisch wie abgebildet aufgenommen worden sein? Enthält das Bild glaubhaft zufällige, störende Elemente?
- Tiefe: Stimmen die Größenverhältnisse von Objekten mit der Perspektive überein? Werden Objekte weiter hinten im Bild kleiner und unschärfer?

VIDEOS

- Verhaltensweisen: Blinzelt die Person? Macht sie kleine Gesten, Gesichtsbewegungen, Augenbewegungen, die zufällig und dadurch menschlich erscheinen?
- Integrität: Schließt das Gesicht immer bündig mit dem Kopf ab, ohne unscharfe Ränder oder Verwischungen und Verschiebungen?
- Dynamik: Verändern sich Licht und Schatten im Gesicht einer Person entsprechend, wenn sie sich bewegt?
- Einheit von Ton und Bild: Stimmen die Lippenbewegungen einer Person mit dem Gesagten überein?
- Optik: Kann ich im Gesicht den gleichen Grad an Details und Schärfe wahrnehmen wie in der Kleidung oder den Haaren? Stimmt die Perspektive? Verhält sich die Schärfentiefe wie erwartet?

TEXT

- Wortwahl: Enthält der Text individuelle Formulierungen und geht auf ein konkretes Thema mit spezifischen Worten ein, ohne durchgehend generell, oberflächlich und diplomatisch formuliert zu sein?
- Orthographie: Enthält der Text ein gewisses erwartetes Maß an Grammatik- und Rechtschreibfehlern, die die meisten Menschen machen?
- Richtigkeit: Enthält der Text korrekte Darstellungen von Gegebenheiten, ohne falsche, aber überzeugend formulierte Angaben?
- Tonalität: Passt der Stil des Textes zur Situation und zur Verfasserin oder zum Verfasser? Ist er so förmlich oder informell, allgemein oder spezifisch wie erwartet?

Lea Uhlenbrock ist wissenschaftliche Mitarbeiterin und Doktorandin der Informatik an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Ihre Forschungsschwerpunkte sind Bildforensik und maschinelles Lernen sowie die Erkennung von KI-generierten Bildern.