Eva Wackenreuther

Auf der Suche nach der Wahrheits-maschine

Niemand will auf Fakes reinfallen. Gleichzeitig wird genau das immer wahrscheinlicher. Kein Wunder also, dass Redaktionen wie Privatpersonen gleichermaßen auf der Suche nach dem Allheilmittel gegen die Epidemie der Falschinfo sind. Oft suchen sie die Lösung bei technischen Werkzeugen. Warum das zu kurz greift und was es braucht, um KI-generierte Katzen und veränderte Videos noch zu erkennen.

Wir alle fürchten den Fake, der als echt durchgeht: dass wir, ohne es zu bemerken, Blödsinn glauben, dass uns eine Meldung mit dubioser Quellenlage untergejubelt wird oder dass uns in einem unaufmerksamen Moment ein manipuliertes Video durchrutscht. Berechtigte Sorgen. Tricksereien in sozialen Medien sind an der Tageordnung, Desinformationskampagnen genauso. Worüber sich viel weniger Menschen Gedanken machen, ist die Kehrseite: das Echte, das als Fake durchgeht.

Davon profitieren vor allem diejenigen, die Misstrauen säen und Medienskepsis ernten. Der Begriff „liar's dividend" beschreibt genau dieses Phänomen. Wenn es darum geht, sich aus der Verantwortung zu stehlen, wischen öffentliche Figuren korrekte Informationen mit dem Verweis auf „Fakes" und Künstliche Intelligenz vom Tisch. Mit dieser Strategie sind sie oft erfolgreich, schließlich werden Deepfakes jeden Tag realistischer und KI- Videos sind mittlerweile täglich in unseren Timelines zu finden.

Skepsis als Stolperfalle

Wenn seriöse Medien Informationen posten, müssen sie ebenfalls gegen diesen Sog an Unsicherheit ankämpfen. Kaum ein Video mehr, unter dem nicht irgendwer „Ist das echt?" kommentiert. Selbst wenn die Information stimmt, nimmt manchmal der Zweifel überhand. Die Nachrichtenseite der „Zeit im Bild" muss darauf immer wieder reagieren. Da gibt es beispielsweise ein Video eines Schweines, das von einem Baggerfahrer aus den Fluten gerettet wird, nachdem es in China zu Überschwemmungen gekommen ist. Das Video wirkt für viele unrealistisch, sie schieben es sofort auf vermeintliche Manipulation. Die „Zeit im Bild" erklärt die Hintergründe.

Brisanter wird das bei politisch aufgeladenen Themen, die nach demselben Muster angezweifelt werden. Als der damalige kanadische Regierungschef Justin Trudeau anlässlich seines Rücktritts einen Sessel aus dem Parlament trug und dabei die Zunge herausstreckte, hielten viele die Aufnahme für einen Fake. Das Verifikationsteam des ORF muss in einem separaten Posting die Echtheit des Fotos prüfen und erklären, wie es zu der Szene kam: In Kanada gibt es die Tradition, dass scheidende Premiers ihren Stuhl, beziehungsweise eine Replik davon, mitnehmen dürfen.





| Bildschirmaufnahme: Zeit im Bild auf Instagram

Die Arbeit von Journalistinnen und Journalisten besteht also nicht mehr bloß darin, Falschinfos aufzudecken und ihre Verbreitung einzudämmen; immer öfter ist sie auch ein Ringen um die Glaubwürdigkeit echter Nachrichten. Ein Fake ist eine Gefahr, genauso aber auch die Sorge vor dem Fake, der uns an Institutionen und Medien zweifeln lässt. Kein Wunder also, dass die Sehnsucht nach Erkennungswerkzeugen steigt.

Kein guter Schiedsrichter

Während KI-Fotos und -Videos manchmal noch tollpatschig aussehen, haben KI-Texte ihren Weg schon in jede Ritze gefunden. Schülerinnen wurden beim Schummeln erwischt, Politiker beim Plagiieren und sogar dem Papst wurde vorgeworfen, in seiner Enzyklika über Künstliche Intelligenz auf KI zurückgegriffen zu haben. Zumindest wenn man KI-Erkennungstools Glauben schenken will und kann. Das ist aber oft alles andere als gewiss.

Was Texte betrifft, gilt das Tool *Pangram* als Goldstandard. In aktuellen Debatten werden dieses oder ähnliche Detektoren gerne als Beweismittel ins Spiel gebracht. Und während die Tools tatsächlich bemerkenswert viel leisten, betonen Fachleute auch, dass sie den großen Versprechen nicht gerecht werden. Das Magazin „The Atlantic“ titelte etwa im Mai: „Amerika hat ein Pangram-Problem“.

Auch wenn solche Erkennungswerkzeuge meist richtig liegen: Im Einzelnen tun sie es keineswegs immer. So stufte das Tool ZeroGPT im Jahr 2023 sogar die US-Verfassung als angeblichen KI-Text ein. Die Erkennungstools raten deshalb im Kleingedruckten meist selbst dazu, akademische Integrität nicht alleine anhand der Einschätzung zu bewerten. Kurz gesagt: Auch die besten Erkennungstools – und es gibt gute – machen Fehler. Wenn darauf aber eine ganze Kette an Meinungen und Entscheidungen fußt, wird das problematisch.

Hinweis nicht Beweis

Nicht nur für Text, auch für Videos und Bilder gibt es unzählige Möglichkeiten, KI-generierte Inhalte zu erstellen – und fast genauso viele, um diesen auf die Schliche zu kommen. Ein konstantes Katz-und-Maus-Spiel, manchmal mit irreführenden Ergebnissen. Das Analyseergebnis solcher Tools ist also erst einmal ein Hinweis, kein endgültiger Beweis.

Da gibt es etwa den Fall von Vizekanzler Andreas Babler: Er veröffentlichte ein Video eines kurzen Statements auf Englisch. Manche meinen: in zu gutem Englisch, das müsse KI sein. Ein oft genutztes und meist





| Bildschirmaufnahme: Zeit im Bild auf Instagram

recht zuverlässiges Detektionstool spuckte ein Ergebnis aus, das diese Einschätzung zu bestätigen schien. Nachforschungen des ORF-Verifikationstools ergaben ein etwas anderes Bild: Im Video kam KI zwar zum Einsatz, aber nicht, um einen Deepfake von Grund auf zu generieren, sondern lediglich um Hall in der Tonspur zu entfernen. Ein üblicher Vorgang bei der Videobearbeitung, der dem Politiker fast viel Glaubwürdigkeit gekostet hätte.

Bei Babler ist genau das geschehen, was immer wieder vorkommt: Ein Ergebnis wird falsch interpretiert. Überhaupt ist es gar nicht so einfach, die Ergebnisse dieser Werkzeuge richtig zu lesen. Für die einen sind 90 Prozent ein fast perfektes Ergebnis, für den nächsten sind 10 Prozent Zweifel bereits massiv. Zudem kommt jedes Tool anders auf die Ergebnisse: Dasselbe Foto erhält von dem einen Werkzeug ein Urteil von 79 Prozent KI-Wahrscheinlichkeit, vom nächsten 86 Prozent. Der Weg dahin bleibt meist eine Blackbox.

Damit gilt auch, dass wir oft nicht merken, wann ein Tool mit einer neuen technischen Entwicklung nicht mehr mithält. Ein Tool, das heute verlässlich funktioniert, kann morgen schon veraltet sein. Darauf sind wir schlecht vorbereitet. Ein KI-Detektor, der fast immer richtig liegt und den man deshalb nicht infrage stellt, ist in gewisser Weise gefährlicher als einer, der oft danebenliegt.

Vermenschlichung der KI-Welt

Dass mittlerweile viele Augen auf KI-Inhalte gerichtet sind, ist auch deren Erstellern nicht verborgen geblieben. Sie stecken teils viel Energie in die Verschleierung der KI-Herkunft, sei es durch manuelle Nachbearbeitung oder aber mithilfe anderer technischer „Geschütze“.

In manchen Fällen tricksen diese Tools dann wiederum ihre Nutzer aus. Die Nachrichtenagentur AFP legte einem solchen Werkzeug einen ungarischen Literaturklassiker aus dem Jahr 1916 vor, also sicher kein KI-Text. Der Detektor zeigte trotzdem einen hohen Anteil Künstlicher Intelligenz an. Wer das ändern möchte, dem bietet das Tool aber praktischerweise gleich eine „Vermenschlichungs-Funktion“ an, für die man zahlen muss. Möglich wird das Geschäft durch die Ambiguität von KI-Erkennungstools.

Dazu kommt, dass die allgegenwärtige Flut von KI-Texten auch unseren natürlichen Sprachgebrauch verändert. So manch einer verzichtet bewusst auf den symbolisch gewordenen Gedankenstrich, um nicht der KI-Schreiberei beschuldigt zu werden. Noch mehr passiert aber unbewusst: In „echte Sprache“ schleichen sich Wendungen ein, die wir ständig lesen. Von außen den

Unterschied zu belegen, ist oft unmöglich.

Verifikation bleibt Handarbeit

Wenn Tools an ihre Grenzen stoßen, greifen die traditionellen Prinzipien von Qualitätsjournalismus. Der Schlüssel liegt dann nicht mehr in einem einzigen Wunderwerkzeug der Verifikation, sondern in einem ganzen Schweizer Taschenmesser, das verschiedene methodische Zugänge kombiniert. Die Basis bilden meist W-Fragen: Wer hat warum ein Video oder Bild veröffentlicht, wann und wo wurde es aufgenommen, was zeigt es? Viele Fakes scheitern nämlich nicht an technischer Raffinesse, sondern an schlampigen Antworten auf genau diese simplen Fragen. Informationen über das Wann, Wo und Wie waren schon vor der KI nötig, um Irreführung zu vermeiden. Schließlich kann ein altes Video in einer neuen Situation wieder kursieren und einen falschen Eindruck erwecken oder ein Foto lediglich eine völlig aus dem Kontext gerissene Momentaufnahme sein.

Wer länger zu den genauen Umständen einer Aufnahme recherchiert, findet oft den goldenen Schlüssel der Verifikation: eine zweite Quelle. Jemand anderer hat ebenfalls sein Smartphone gezückt, um eine denkwürdige Szene einzufangen, und Details aus den Videos decken sich.

Manches lässt sich aber nicht nur am Bildschirm klären. Für manch einen Ort gibt es keine Straßenansicht im Internet, man muss zum Hörer greifen und nachfragen. Hilfreich können Augenzeugen oder Expertinnen sein. Wer mit einer Situation vertraut ist, dem fallen oft ganz andere Dinge auf: „Das Video kann nicht vom besagten Abend sein, wir hatten den ganzen Abend Hunger, aber auf diesem Tisch stehen Brötchen."

Um solche Nachforschungen anzustellen, braucht es Zeit und Ressourcen. Angesichts schneller Sehgewohnheiten, Unmengen an Videos, besser werdender Fakes und personell ausgedünnter Redaktionen ist das eine Mammutaufgabe. Qualitätsjournalismus ist dafür ein wichtiger Grundpfeiler, wirklich lösen wird das „Wahrheitsproblem" aber nur eine medienkompetente Gesellschaft.

Eva Wackenreuther prüft beim Österreichischen Rundfunk für das Verifikationsressort „ORF defacto" die Echtheit von Videomaterial und berichtet über Desinformation und Digitalphänomene.



Entgeltliche Einschaltung