



Kapitel 3: Fakes und Falschinfos erkennen

Wenn die KI lügt, wenn Menschen lügen und wie du beides merkst.

- 1 Warum KI überzeugend lügt
- 2 Quellen wirklich prüfen
- 3 Scam, Fake-Shops, Phishing
- 4 Bilder, Videos, Stimmen
- 5 Meisterstück: eine Behauptung wirklich prüfen

Lektion 1 von 5

Warum KI überzeugend lügt

Ziel: Ich erkenne die Stellen, an denen eine KI am wahrscheinlichsten erfindet.

Die KI lügt nicht, sie schätzt

Eine KI hat kein Gefühl dafür, was sie weiß und was nicht. Sie erzeugt in beiden Fällen denselben souveränen Ton.

Der Fachbegriff dafür ist Halluzination: eine erfundene Angabe im Stil einer belegten. „Das weiß ich nicht“ ist für ein Sprachmodell die unwahrscheinlichste Fortsetzung.

Merksatz

Souveräner Ton sagt gar nichts. Prüf die Angaben, nicht das Gefühl.

Die Risikoliste

Je genauer eine Angabe wirkt, desto eher ist sie erfunden. Der Grund: Das Format einer Quellenangabe erzeugt ein Modell genauso mühelos wie ihren Inhalt.

- Quellen, Studien, Buchtitel, Seitenzahlen
- Paragraphen, Fristen, Aktenzeichen
- Zahlen mit Komma und Jahresangaben
- Zitate von echten Personen

Wo die Erfindungen am dichtesten sitzen

Halluzinationen verteilen sich nicht gleichmäßig über eine Antwort. Sie sitzen dort, wo eine genaue Form erwartet wird und der Inhalt selten ist.

Das ist der Grund, warum ausgerechnet die Stellen erfunden werden, die am meisten nach Beleg aussehen: Eine Jahreszahl mit vier Ziffern und ein Autorenname mit Initiale sehen aus wie Beleg und sind reine Formfüllung.

- Zahlen: Prozente, Preise, Fristen, Seitenzahlen
- Eigennamen: Autoren, Firmen, Studentitel
- Verweise: Paragraphen, DOI, Internetadressen
- Wörtliche Zitate mit Anführungszeichen

Lektion 2 von 5

Quellen wirklich prüfen

Ziel: Ich prüfe, ob eine Quelle existiert, die Behauptung trägt und für die Aufgabe geeignet ist.

Drei getrennte Prüfungen

Prüf bei einer Quellenangabe zuerst die Existenz, dann den Inhalt und schließlich die Eignung für deine Aufgabe. Der Zeitbedarf hängt von Quelle und Zugang ab: Manches ist sofort klar, anderes verlangt Katalogsuche oder das Lesen des Volltexts.

Wenn eine schwer zugängliche oder unklare Quelle länger dauert, ist das kein persönliches Versagen. Lass sie offen oder hol Hilfe, statt die Lücke mit einer Vermutung zu füllen.

- Existenz: Titel, Autor, DOI, Verlag und Bibliothekskatalog abgleichen
- Inhalt: Steht im Abstract oder Volltext wirklich, was behauptet wird?
- Eignung: Passt die Quelle zu Frage, Zeitraum und benötigter Belegart?

Merksatz

Erst nachsehen, ob es die Quelle gibt. Dann, wer dahintersteht. Dann, ob sie das wirklich sagt.

Woran eine Quelle greifbar wird

Autor, Titel und Jahr reichen nicht: Diese drei lassen sich zusammensetzen. Greifbar wird eine Quelle erst durch eine Kennung, die auf genau ein Werk zeigt.

Bei Aufsätzen ist das die DOI, bei Büchern die 13-stellige ISBN, bei Webseiten die vollständige Adresse mit Abrufdatum. Ohne eine davon hast du eine Beschreibung, keine Fundstelle.

Merksatz

Verlang die Kennung ausdrücklich mit: Nenn zu jeder Quelle die DOI oder die ISBN, und schreib „ohne Kennung“ dahinter, wenn keine vorliegt.

Lektion 3 von 5

Scam, Fake-Shops, Phishing

Ziel: Ich erkenne Betrugsmerkmale in einer Nachricht, ohne auf einen Link zu klicken.

Immer dieselben vier Muster

Betrug ist erschreckend unkreativ. Wer die vier Muster kennt, erkennt die meisten Fälle in Sekunden.

- Zeitdruck: „innerhalb von 24 Stunden“
- Ungewöhnlicher Zahlweg: Gutscheinkarten, Krypto, Freunde-Überweisung
- Absender fast richtig: dhl-paket-service.net statt dhl.de
- Etwas, das du nicht erwartet hast: Paket, Gewinn, Erbe, Job

Was sich durch KI wirklich geändert hat

Die alten Erkennungsmerkmale waren sprachlich: schlechte Grammatik, seltsame Anrede, gebrochenes Deutsch. Die sind weg. Ein Sprachmodell schreibt fehlerfrei und in jedem Ton.

Was bleibt, sind die Merkmale, die am Vorgang hängen und nicht am Text: Absenderadresse, Zahlungsweg, Zeitdruck, Kanalwechsel. Sie ändern sich nicht, weil sie zum Betrug gehören und nicht zur Sprache.

Merksatz

Prüf nie den Text, sondern den Weg: Wer schickt es, wohin soll das Geld, warum eilt es, und warum soll das Gespräch woanders weitergehen?

Was du niemals tust

Keine Links aus verdächtigen Nachrichten anklicken, auch nicht „nur zum Schauen“. Der sichtbare Linktext ist frei wählbar und sagt nichts über das Ziel.

Willst du eine Zieladresse analysieren, übernimm höchstens die Domain oder schreibe die Adresse mit getrennten Punkten als nicht anklickbare Zeichenkette ab. Fehlt die Zieladresse, ist das Linkziel ausdrücklich nicht beurteilbar. Zum Konto gehst du nur über eine selbst gefundene offizielle Adresse.

Merksatz

Nie den Link. Immer selbst tippen.

Lektion 4 von 5

Bilder, Videos, Stimmen

Ziel: Ich weiß, woran man erzeugte Bilder erkennt und warum das immer schwerer wird.

Die alten Merkmale verschwinden

Sechs Finger und verbogene Brillen waren gestern. Bei aktuellen Modellen sind Hände und Gesichter meist fehlerfrei.

Was bleibt: Zusammenhänge. Schrift im Hintergrund, Spiegelungen, Schmuck, der die Seite wechselt, Schatten aus zwei Richtungen.

- Text im Bild: Schilder, Etiketten, Bücher
- Spiegel, Fensterscheiben, Pupillen
- Ohren und Schmuck bei bewegten Personen
- Übergänge zwischen Haar und Hintergrund

Merksatz

Nicht mehr auf Finger achten, sondern auf Zusammenhänge: Schrift, Spiegelungen, Schatten.

Herkunftsdaten statt Bildmerkmale

Weil man synthetischen Bildern die Herkunft immer weniger ansieht, verschiebt sich die Prüfung auf die Datei: C2PA ist ein Industriestandard, der Kamera, Bearbeitungsschritte und Erzeugung in drei signierten Angaben mitführt.

Die Einschränkung ist erheblich: Die Angaben gehen beim Hochladen in soziale Netze meist verloren, und ihr Fehlen beweist nichts. Vorhandene und geprüfte Herkunftsdaten sind ein Beleg, fehlende sind keiner.

Merksatz

Kein einzelnes Merkmal entscheidet. Was entscheidet, ist die Übereinstimmung mehrerer unabhängiger Wege: Herkunftsdaten, Rückwärtssuche, zweite Quelle.

Meisterstück: eine Behauptung wirklich prüfen

Ziel: Ich prüfe eine Behauptung so, dass ich am Ende sagen kann, worauf sie beruht und wo sie endet.

Was du hier machst

Du nimmst eine Behauptung, die dir in den letzten Wochen wirklich begegnet ist, und prüfst sie zu Ende. Nicht bis zum Gefühl, sondern bis zu einem Satz, der sagt, worauf sie beruht.

Das dauert 20 bis 30 Minuten, und der längste Teil ist nicht die Suche, sondern das Zerlegen: Die meisten Behauptungen bestehen aus drei bis vier Einzelaussagen, von denen zwei stimmen.

Merksatz

Das Ergebnis ist selten „stimmt“ oder „stimmt nicht“. Es ist meistens: „stimmt unter dieser Bedingung, und die stand nicht dabei“.

Das nimmst du mit

- Ich erkenne die Stellen, an denen eine KI am wahrscheinlichsten erfindet.
- Ich prüfe, ob eine Quelle existiert, die Behauptung trägt und für die Aufgabe geeignet ist.
- Ich erkenne Betrugsmerkmale in einer Nachricht, ohne auf einen Link zu klicken.
- Ich weiß, woran man erzeugte Bilder erkennt und warum das immer schwerer wird.
- Ich prüfe eine Behauptung so, dass ich am Ende sagen kann, worauf sie beruht und wo sie endet.

Welche Prüfung liefert bei einem verdächtigen Bild die stärkste Einordnung?

- A** Eine Rückwärtssuche: Findet sie nichts, ist das Bild neu oder mit KI erzeugt
- B** Eine Rückwärtssuche: Ihr frühester Treffer ist die Originalquelle des Bildes
- C** Die Rückwärtssuche als Indiz, dazu Quelle, Kontext, Ort und eine unabhängige Bestätigung

Auflösung

Richtig ist C: Die Rückwärtssuche als Indiz, dazu Quelle, Kontext, Ort und eine unabhängige Bestätigung

Eine Suchmaschine zeigt früh auffindbare oder ähnliche Verwendungen, nicht automatisch den Ursprung. Fehlende Treffer beweisen weder, dass ein Bild neu ist, noch dass es KI-erzeugt ist. Tragfähig wird die Einordnung erst mit Originalquelle, Veröffentlichungskontext, Ort und unabhängiger Bestätigung.