33 Minuten für…

Coffee Lectures

„KI generierte Inhalte erkennen“

Claudia Piesche

Universität zu Köln :: Universitäts- und Stadtbibliothek

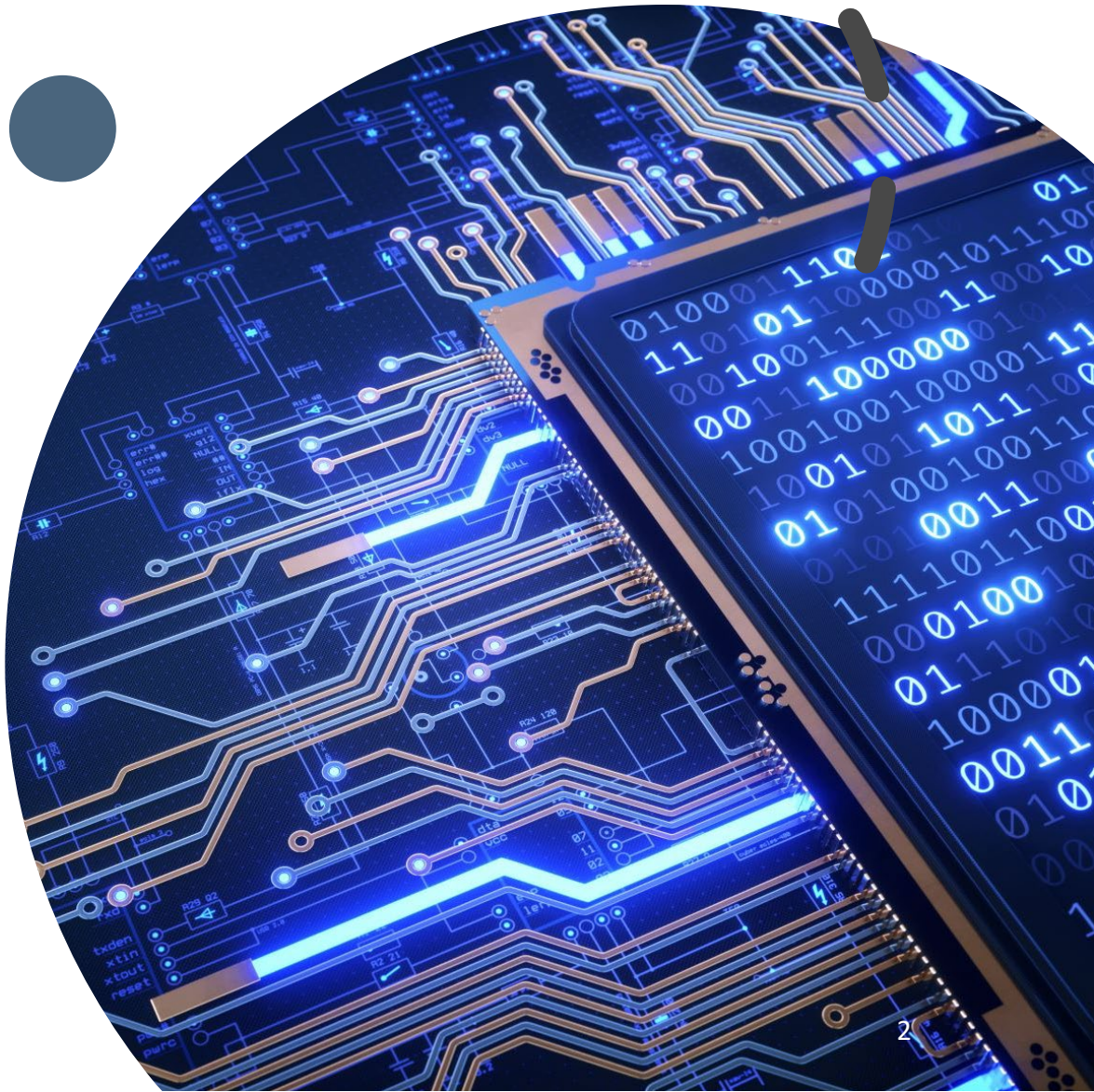
IT-Dienste

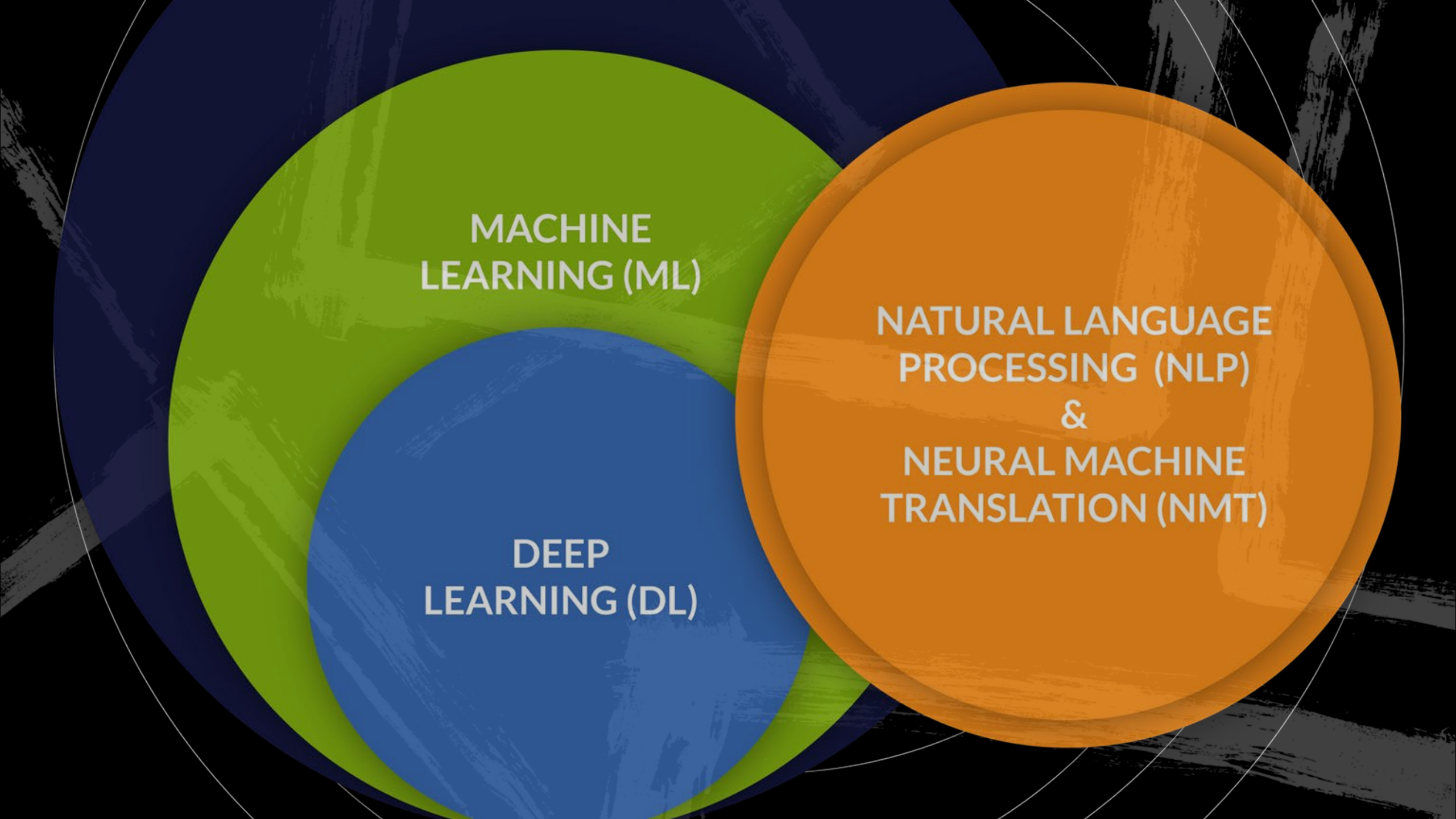
Universitäts- und Stadt Bibliothek Köln

Künstliche Intelligenz

Fähigkeit von Computern und Maschinen, Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern:

→ Schlussfolgern, Problemlösen, Verstehen natürlicher Sprache und Erkennen von Mustern




MACHINE
LEARNING (ML)
DEEP
LEARNING (DL)
NATURAL LANGUAGE
PROCESSING (NLP)
&
NEURAL MACHINE
TRANSLATION (NMT)

KI-generierten Inhalt erkennen?

Die Bilder wurden mithilfe des Dall-E-Modells generiert, das von OpenAI entwickelt wurde. Dall-E ist ein neuronales Netzwerk, das Bilder aus Textbeschreibungen erzeugt. Die Textbeschreibungen zur Erzeugung der Bilder wurden von Mark Eschweiler erstellt.

Erkennung KI-generierter Texte

Merkmalsbasierte Ansätze

- Häufigkeit von Begriffen
- Sprachgewandtheit oder Lesbarkeit → bei längeren Texten haben KIs Probleme mit Kohärenz und klarer Struktur
- Komplexe Phrasen → bestimmte idiomatische Redewendungen nicht in KI-Texten
- Grundlegende Textmerkmale → Muster bei Zeichensetzung oder Länge von Sätzen / Absätzen

Erkennung KI-generierter Texte

Neuronale Sprachmodellansätze

- „Modelle erkennen sich selbst“ → Nutzung neuronaler Netze zum Vergleich KI-generierter und menschlicher Texte
- Klassifikation anhand Merkmalsvektoren
- Methoden: Zero-shot Approach, Fine-tuning Approach

Erkennung KI-generierter Texte

Menschengestützte Methoden

- statistischer oder neuronaler Ansatz in Kombination mit einem menschlichen Analysten zur Überprüfung
- Visualisierung automatischer Erkennung + menschliche Mustererkennung
- ungeschulte Nutzer:innen → entspricht Zufallserkennung
- Geschulte Nutzer:innen erkennen KI-Texte zuverlässiger
- menschliche Urteile werden durch intuitive, aber fehlerhafte Heuristiken beeinflusst → Pronomen: Ich / Wir, Bindewörter oder Abkürzungen

Tools zur Erkennung von ChatGPT-Texten

GPT-2 Output Detector Demo

- neuronales Netz als Klassifizierer → lernt wie typisch maschineller Text aussieht und wie ein typisch menschlicher

Giant Language Model Test Room

- Software berechnet Wort für Wort, mit welcher Wahrscheinlichkeit das jeweils nächste Wort von dem zugehörigen Sprachmodell ergänzt werden würde
- Wörter mit einer hohen Wahrscheinlichkeit werden grün eingefärbt, unwahrscheinliche Wörter rot und sehr seltene Wörter violett

Tools zur Erkennung von ChatGPT-Texten

DetectGPT

- arbeitet auf ganzen Sätzen
- berechnet Wahrscheinlichkeit mit der ein Sprachmodell den zu prüfenden Satz erzeugen würde
- inhaltsgleiche Umformulierung des Satzes + Berechnung der Wahrscheinlichkeiten
- Wahrscheinlichkeit des ursprünglichen Satzes > Wahrscheinlichkeiten der geänderten Sätze → Produkt eines Sprachmodells

Tools zur Erkennung von ChatGPT-Texten

GPTZero

- berechnet die „Perplexity“ für ein Stück Text - wird in der NLP verwendet, um die Güte eines Sprachmodells zu bestimmen
- Perplexity ist ein Wert, der ausdrückt, wie überraschend das nächste Wort in einem Text ist

Zukunft? Unsichtbares Wasserzeichen

- Wasserzeichen-Software erzeugt eine Liste von Wörtern, die das Sprachmodell nur mit einer verringerten Wahrscheinlichkeit wählen darf
- Parameter für die Erzeugung dieser Liste werden mit dem Text veröffentlicht

Tools zur Erkennung von ChatGPT-Texten

Classifier

- Aufklärungsquote liegt bei englischsprachigen Texten bei 26%
- benötigt mindestens 1.000 Zeichen
- unterscheidet zwischen „very unlikely, unlikely, unclear if it is, possibly, or likely AI-generated“

Originality.AI

- bisher fortschrittlichste Tool zur Erkennung von KI-generierten Texten
- Nur Bezahlversion verfügbar

Intellektuelle Erkennung / typische Merkmale

- Keine Phrasen oder außergewöhnliche Wortkombinationen
- keine orthographischen Fehler
- Keine Wortzusammensetzungen oder Neologismen (Wortneuschöpfungen)
- Wiederholungen, wenig sprachliche Varianz
- Nutzung vieler Keywords

KI-Texte erkennen – Linkliste der GEW

- KI Texte erkennen - Diese 12 Tools habe ich getestet (jens.marketing)
- KI-Text erkennen in 2023: 14 Detektoren ausführlich getestet (blogmojo.de)
- KI Texte erkennen: Die besten KI Text Detektoren 2023 (kopfundstift.de)
- KI Texte erkennen 2023: 6 Besten KI Text Detektoren, Tools, Prüfer (onlinemarketing-mastermind.de)
- Der ChatGPT-Guide für Lehrkräfte — Manuel Flick

KI-generierte Bilder erkennen?

Tipps zum Erkennen KI-generierter Bilder

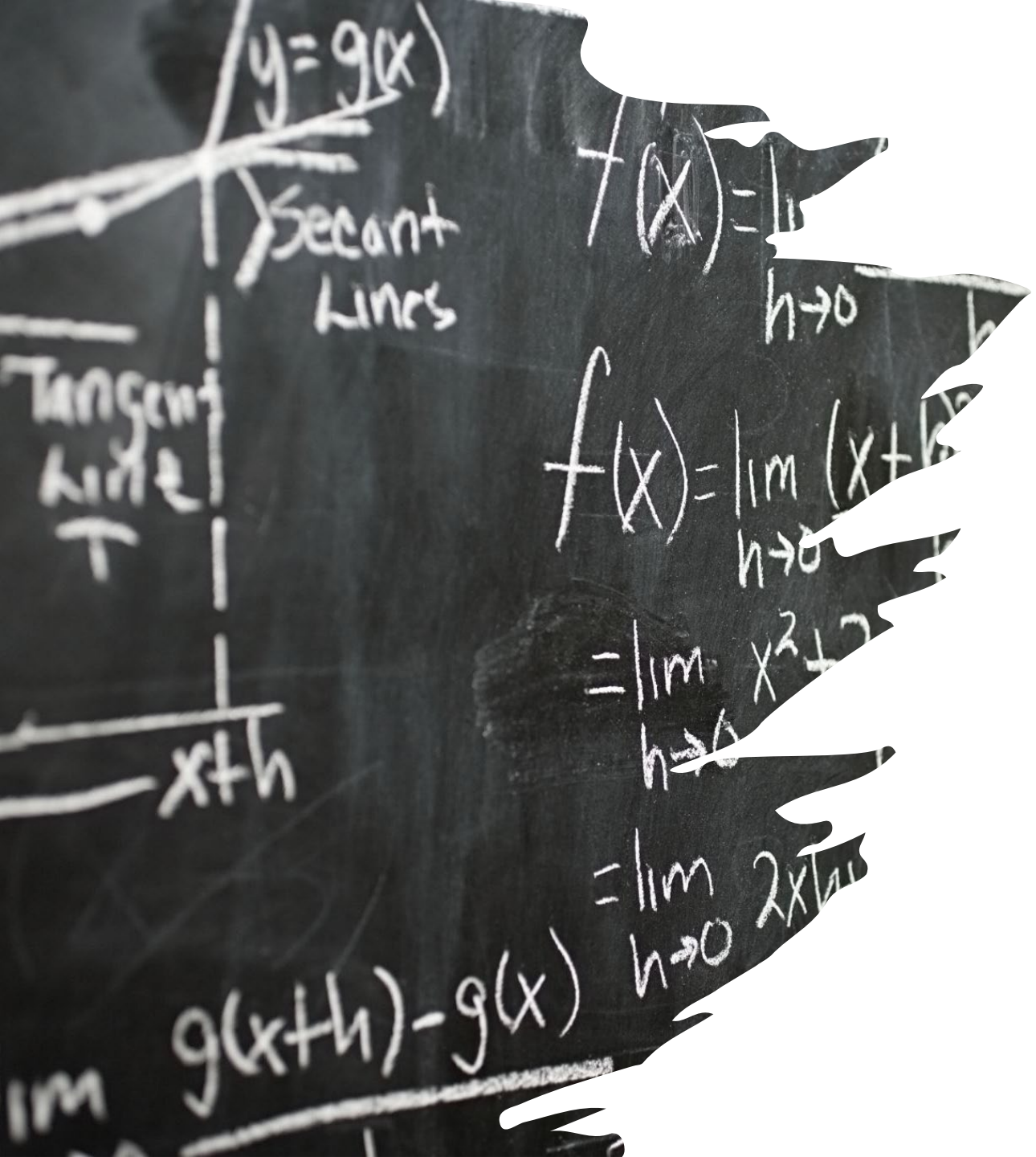
- Heranzoomen und genau untersuchen → hoch auflösenden Versionen des Bildes nutzen und auf Details zoomen
- Quelle / Herkunft des Bildes recherchieren → Kommentare, Bilderrückwärtssuche
- Stimmen alle Körper-Proportionen der abgebildeten Personen?
- typische KI-Fehler bei Bilddetails: Hände, Ohren, Finger, Brille, Schmuck
- Wirkt das Bild künstlich und geglättet? Kann ein so perfektes, ästhetisches Bild mit makellosen Menschen wirklich echt sein?
- Bildhintergrund → unscharf, deformiert, Wiederholungen?

Tools zur Verifikation

- Browser-Plugin zum Verifizieren von Bildern: InVid & WeVerify (https://www.invid-project.eu/tools-and-services/invid-verification-plugin/)

Erkennen bearbeiteter Bilder: Noise Analysis (Pixeldichte), Error Level Analysis (JPEG-Komprimierungsrate)

- Forensically
- Foto Forensics



Fazit

KI-generierte Inhalte erkennt man am besten durch Hinterfragen, Recherchieren und Plausibilitätschecks

Tools können nur Anhaltspunkte liefern, die Interpretation muss durch den informierten Nutzenden erfolgen.

In diesem Sinne, viel Vergnügen beim Ausprobieren:

https://newsevaluator.com/

https://www.getbadnews.de