(Wie) können Menschen echte und KI-generierte Bilder unterscheiden?

Semesterprojekt im Master IDMM | Seminar Forschungskolloquium, Prof. Dr. phil. Michael Meng (WS 2023/24)
Autor*innen: Veronika Erl, Jonas Pfennig, Julian Wucherer, und Luise Winkler

Einführung

Angesichts der schnellen Fortschritte in der Bildgenerierung durch künstliche Intelligenz (KI) gewinnen Methoden und Fähigkeiten zum Erkennen und Überprüfen der Echtheit bzw. Manipulation von Bildinhalten immer mehr Bedeutung. Diese Studie kombiniert die Methoden **Eyetracking**, d.h. das Aufzeichnen von Blickbewegungen, und **qualitative Interviews**, um die Wahrnehmungsprozesse beim Unterscheiden KI-generierter und echter Bilder zu untersuchen. Konkret wurde überprüft, ob die Teilnehmenden (TN) der Studie in der Lage sind, **KI-generierte Bilder zu erkennen,** und welche Faktoren dabei wichtig sind – insbesondere, inwiefern **bestimmte Merkmale oder Bereiche in diesen Bildern** für eine korrekte Einordnung relevant sind.

Forschungsstand

- Aktuelle Studien mit Fokus auf menschlichen **Portraits**: TN kaum bzw. nicht in der Lage, KI-generierte von echten Bildern zu unterscheiden (Bray et al. 2023, Nightingale & Farid 2022, Shen et al. 2021, Tucciarelli et al. 2022)
- Frage nach der **Existenz gemeinsamer Merkmale KI-generierter Bilder**, die von den TN zur Unterscheidung genutzt werden (Miller et al. 2023)
- Unterscheidungsmerkmale in einigen Studien als **Hilfestellung** für TN genutzt, um Unterscheidung zu erleichtern (Bray et al. 2023, Nightingale & Farid 2022), jedoch kaum systematisch erforscht (Nightingale & Wade 2021), **schwer greifbar** durch:
  - » Schnelle Weiterentwicklung von KI-Bildgeneratoren (West & Bergstrom 2019)
  - » Komplexe und subjektive Bildwahrnehmungsprozesse, oft intuitiv („gut feeling“)
- **Eyetracking-Studie** zu „gaze area“ (betrachtete Bildfläche) und „gaze spread“ (Längenmaß für Sakkaden): Korrelation dieser Maße mit Korrektheit (Caporusso et al. 2020)

Methode

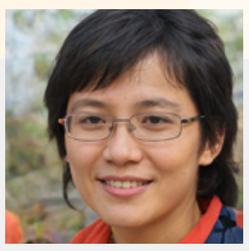
Teilnehmende
- 14 Testpersonen (6 Frauen, 8 Männer)
- 25 – 37 Jahre alt (ø 30,7 Jahre)
- Großteil im Medienbereich tätig

Variablen
- Unabhängig: Bilder (echt/KI)
- Abhängig: Korrektheit der Antwort, Sakkadenlänge & Fixationsdauer/-ort
- Kontrollvariablen: Selbsteinschätzung, Einstellung zu KI & Vorerfahrung der TN

1) Leitfadeninterview
- Vorbesprechung zum Erfassen von Alter, Beruf, Mediennutzung, Erfahrungen mit Bildbearbeitung & KI-Bildern
- Selbsteinschätzung der eigenen Fähigkeit zum Erkennen der KI-generierten Bilder sowie Einstellung gegenüber KI

2) Eyetracking mit BeGaze
- 30 Bilder (15 echt, 15 KI-generiert) einzeln für max. 20 Sekunden gezeigt
- 3 Bildkategorien mit je 10 Bildern in randomisierter Reihenfolge
- Danach Entscheidung: KI-Bild oder nicht? (ja/nein)

3) Leitfadeninterview
- Nachbesprechung, wobei Bilder nach Kategorien erneut gezeigt wurden
- Fragen zur Entscheidungsfindung und auffälligen Bildmerkmalen/-bereichen



1) Portraits
Menschliche Gesichter in Nahaufnahme

2) Menschen
Personen in größerem Bildausschnitt

3) Objekte
Unbelebte Gegenstände wie Essen

Hypothesen

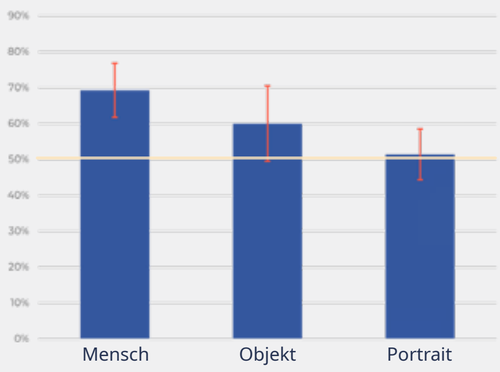
1 Die TN können KI-generierte und echte Bilder unterscheiden und **ordnen diese korrekt zu,** d.h. sie raten nicht nur.

2 Zwischen TN mit korrekten und jenen mit inkorrekten Antworten gibt es Unterschiede bei a) den **Sakkadenlängen** und b) **Dauer & Ort der Fixationen**.

3 Die TN nennen bestimmte gemeinsame **Merkmale der KI-generierten Bilder**, die zur Unterscheidung wichtig sind.

Ergebnisse

1 Mittelwert des Anteils korrekter Antworten: **60,2% (± 5,8%)**
- Signifikant mehr als 50%: Nullhypothese „Raten“ verworfen (γ = 95%)
- Auswertung nach Kategorien:
  - » Korrektes Zuordnen bei Menschen (p = 0,0002) und Objekten (p = 0,08)
  - » Raten bei Portraits (p = 0,69)
- Unterschiedliche Leistung je nach TN (40 – 76,7%) und Bild (28,6 – 100%)



Korrektheit der Antworten nach Bildkategorie

2 **a)** Keine Korrelation zwischen Länge der Sakkaden & Korrektheit je TN
- Bei Sakkadenamplitude & Korrektheit je Bild schwach negativ lineare Korrelation (r = -0,37, p = 0,04)

b) Grafischer Vergleich von Heatmaps mit den längsten Fixationen der TN mit korrekten und jener mit inkorrekten Antworten: **Unterschiede feststellbar**

falsche Antworten (echtes Foto)

korrekte Antworten (KI-Bilder)

falsche Antworten (echtes Foto)

korrekte Antworten (KI-Bilder)

3 Inhaltsanalyse der Leitfadeninterviews: Identifikation von **drei Arten von Merkmalen** zur Unterscheidung

Stilistische Merkmale
- **Farben:** Sättigung, Farbunterschiede, unnatürliche Farbwirkung
- **glatte Oberflächen:** u.a. Hautbild oder Hintergrund zu glatt
- **Licht:** Lichtverhältnisse unnatürlich
- **Konturen:** Bild(un-)schärfe, Übergang Vordergrund zu Hintergrund

KI-typische Bildfehler
- **Motiv Mensch:** Haare, Zähne, Hände
- **Verstoß gegen Physik:** Schatten, Reflexionen, Spiegelungen, Objektübergänge
- **Details:** (inkongruente) Muster
- **Hintergrund**

Kultureller Kontext
- **Ausdruck von Emotionen**
- **Objektbeziehungen:** gestellte Szene, Pose, Anzahl an Menschen
- **Schönheitsideale**

Diskussion & Interpretation

1 Erkennen KI-generierter Bilder ist für TN teilweise möglich, aber mit Einschränkungen (vgl. vorherige Studien)

2 **a)** Korrelation von Sakkadenlänge und korrekten Antworten ist kaum vorhanden bzw. wenig aussagekräftig
- Leichte Tendenz zu mehr richtigen Antworten bei geringerer Sakkadenlänge
- Widerspruch zu Caporusso et al. (2020)

b) Unterschiede nach Bildkategorie:
- **Mensch**: bei falschen Antworten Fixationen v.a. auf Gesichter – bei korrekten Antworten Fokus auf Hände/Details
- **Objekt**: bei korrekten Antworten oft nur Fixationen auf relevantem Merkmal/Fehler (z.B. Kuchengabel)

Schlussfolgerungen:
- » Finden und **korrektes Interpretieren typischer Merkmale** der Bilder ist wichtig bei der Unterscheidung
- » Vermutlich **Abbruch der Suche**, sobald Merkmal erkannt wurde (These gestützt durch kürzere Scanpaths bei korrekt erkannten KI-Bildern)

3 Unterschiede nach Bildkategorie auch bei den Arten der Merkmale:
- Bei **Objekten** weniger KI-typische Bildfehler als bei Menschen, dadurch verstärkter Fokus auf Stil und Kontext
- Von TN genannte Merkmale stimmen mit Fixationen in Heatmaps überein
- Bewusstsein für Merkmale garantiert nicht deren korrekte Interpretation

Fazit & Ausblick

Die Kombination der Methoden Eyetracking und qualitatives Interview zeigte sich als vielversprechende Möglichkeit, die komplexen Wahrnehmungsprozesse bei der Bildbetrachtung zu untersuchen. Dennoch müssen die Ergebnisse dieser Studie mit dem Vorbehalt betrachtet werden, dass die Stichprobengröße in Zukunft für eine stärkere Aussagekraft erweitert werden sollte. Durch die Auswahl der Testpersonen im privaten Umfeld der Autor*innen ist zudem eine Verzerrung nicht auszuschließen, da Medienberufe und Vorerfahrungen mit Bildbearbeitung etwas überrepräsentiert waren. In zukünftigen Studien könnte die Rolle solcher Vorerfahrungen oder der eigenen Einschätzung der Leistung als Kontrollvariablen stärker berücksichtigt werden.

Literatur

Bray, S.D., Johnson, S.D., Kleinberg, B. (2023). “Testing human ability to detect ‘deepfake’ images of human faces.” Journal of Cybersecurity, Vol. 9, Issue 1, 2023, https://doi.org/10.1093/cybsec/tyad011.

Caporusso, N., Zhang, K., Carlson, G. (2020). “Using Eye-tracking to Study the Authenticity of Images Produced by Generative Adversarial Networks”. 1-6. https://ieeexplore.ieee.org/abstract/document/9179472.

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., Dawel, A. (2023). “AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones.” Psychological Science, 0(0). https://doi.org/10.1177/09567976231207095.

Nightingale, S., Farid, H. (2022). “AI-synthesized faces are indistinguishable from real faces and more trustworthy.” Proceedings of the National Academy of Sciences, USA, 119(8), Article e2120481119. https://doi.org/10.1073/pnas.2120481119.

Nightingale, S., Wade, K. (2022). “Identifying and minimising the impact of fake visual media: Current and future directions.” Memory, Mind & Media, 1, 1-13. Article e15. https://doi.org/10.1017/mem.2022.8.

Shen B., RichardWebster B., O’Toole A., Bowyer K., Scheirer W. J. (2021). A study of the human perception of synthetic faces. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1–8). https://doi.org/10.1109/FG52635.2021.9667066.

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., O’Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. New Media & Society, 21(2), 438-463. https://doi.org/10.1177/1461444818799526.

Tucciarelli R., Vehar N., Chandaria S., Tsakiris M. (2022). On the realness of people who do not exist: The social processing of artificial faces. IScience, 25(12), Article 105441. https://doi.org/10.1016/j.isci.2022.105441.

West, J., Bergstrom, C. (2019). “Which Face Is Real? Learn to spot fake faces at a glance”. Calling Bullshit project, U of Washington. https://www.whichfaceisreal.com/learn.html.

Bildquellen

Portraits: Nightingale, S.J. (2022). “Distinguishing between Real and Synthetic Faces.” Daten verfügbar auf OSF.

Menschen:
Frau auf Festival: © Yuri Arcurs Peopleimages, Freepik.
Kinder auf Spielplatz: © pressfoto, Freepik.
Mann mit Brille: © Freepik, Freepik.
Frau mit Gitarre: © Юлия Клюева, Adobe Stock.
Älterer Mann am Schreibtisch: © Nejron Photo, Adobe Stock.

KI-Bilder:
Midjourney (Frau mit Gitarre, Künstlerin), Playground AI (Kinder) & Leonardo AI (Mann am Schreibtisch & Radfahrer)

Objekte:
Küche: © shangarey, Freepik.
Kuscheltier: © Freepik, Freepik.
Frikadellen: © Michael Algert, foodish cooking.
Plätzchen: © Corinna Eichberger Renneisen, Gustl Magazin.
Kuchen: © New Africa, Adobe Stock.

KI-Bilder:
Leonardo AI (Weihnachten, Frikadelle), Lexica Art AI (Kuscheltier, Küche) & Playground AI (Schokokuchen)

HOCHSCHULE MERSEBURG