Eine empirische Untersuchung zur Erkennung und ethischen Einschätzung KI-generierter Bilder sowie zur Erstellung und Verbreitung sensibler Inhalte über KI-basierte Bildgenerierungstools

Leon Hobelmann

Hochschule für Technik und Wirtschaft Berlin
Wirtschaftsinformatik
Treskowallee 8
10318 Berlin
E-Mail:
mail@leon-hobelmann.de

Birte Malzahn

Hochschule für Technik und Wirtschaft Berlin
Wirtschaftsinformatik
Treskowallee 8
10318 Berli0n
E-Mail:
birte.malzahn@htw-berlin.de

Schlüsselwörter

Digitale Ethik, Midjourney, KI-Bildgenerierung, Desinformation, Plattformrichtlinien

Zusammenfassung

KI-basierte Bildgeneratoren wie Midjourney haben sich als feste Werkzeuge in Wirtschaft und Bildung etabliert. Zu den Vorteilen zählen die Unterstützung von Lern- und Kreativprozessen sowie die Erweiterung gestalterischer Möglichkeiten (Bendel 2025). Mit der zunehmenden Verbreitung dieser Werkzeuge rücken jedoch auch ethische Fragestellungen in den Fokus: KI-generierte Bilder können zur Verbreitung von Desinformation und zur gezielten Beeinflussung öffentlicher Wahrnehmung eingesetzt werden. Des Weiteren kann bei der Generierung von KI-Bildern das Urheberrecht verletzt werden, wenn zum Training der KI geschützte Inhalte verwendet werden (Bird et al. 2023). Plattformrichtlinien und technische Filter adressieren diese Risiken nur begrenzt, da problematische Inhalte trotz Beschränkungen erzeugt oder bzw. untersagte Inhalte mittels Umgehungsstrategien realisiert werden können (Leow 2023).

Vor diesem Hintergrund untersuchte diese Bachelorarbeit empirisch die Erkennung und ethische Bewertung KI-generierter Bildinhalte. Des Weiteren wurde anhand von verfügbaren KI-generierten Bildern auf der Plattform Midjourney analysiert, inwieweit Verstöße gegen die eigenen Richtlinien der Plattform vorliegen (Hobelmann 2025). Der theoretische Teil der Arbeit berücksichtigte KI-ethische Prinzipien wie Transparenz, Verantwortung, Nichtschadensgebot und Fairness, die in KI-Ethikleitlinien als zentrale normative Bezugspunkte hervorgehoben werden (Jobin et al. 2019). Ergänzend wurde der EU AI Act als rechtlicher Rahmen herangezogen, der Transparenzpflichten für KI-Systeme und ein risikobasiertes Regulierungskonzept vorsieht (Europäische Union 2024).

Die empirische Untersuchung wurde mithilfe eines Online-Fragebogens durchgeführt. Der Fragebogen wurde am 28.06.2025 veröffentlicht und im eigenen Umfeld verteilt. Insgesamt wurden 87 verwertbare Datensätze ausgewertet. Die Stichprobe umfasste 52 männliche, 24 weibliche und eine diverse Person; die 23- bis 27-jährigen bildeten mit etwa 33 % die größte Altersgruppe. In einem Three-Alternative-Forced-Choice-Design wurden Bildersets vorgelegt, die jeweils aus drei sehr ähnlichen Darstellungen bestanden, von denen jeweils eine reale Fotografie und zwei mit Midjourney selbst-generierte Bilder waren. Die Teilnehmenden sollten in jedem Set die reale Fotografie auswählen. Ziel war es, die Erkennungsfähigkeit synthetischer Bilder zu erfassen.

Abbildung 1 - Beispiel aus der 3 AFC Aufgabe (Quelle linkes Bild: Adbullah, 2022)

Anschließend bewerteten die Befragten auf Likert-Skalen die ethische Vertretbarkeit von KI-generierten Produktdarstellungen, Karikaturen (z. B. Queen Elisabeth II

auf einem Skateboard), Vorher-Nachher-Bilder (z. B. eine zweiteilige Frontalaufnahme derselben Person: höherer Körperfettanteil vs. deutlich erhöhte Muskelmasse) sowie von Motiven mit potenziell sensiblen Inhalten (konfliktbezogene bzw. körpernahe Darstellungen). Die Ergebnisse zeigen, dass die Teilnehmenden das reale Foto in den Bildersets im Mittel nur in etwa 47 % der Fälle korrekt identifizierten und damit zwar signifikant über dem Zufallsniveau von 33,3 %, aber deutlich unter einer zuverlässigen Erkennungsleistung liegen; signifikante Unterschiede zwischen Altersgruppen und Geschlechtern traten hierbei nicht auf. In der anschließenden Bewertung der Bildkategorien wurden KI-generierte Produktdarstellungen überwiegend als ethisch vertretbar eingestuft. Motive mit Gewalt- bzw. erotischem Bezug erhielten die niedrigsten Zustimmungswerte. Über alle Kategorien hinweg bewerteten männliche Teilnehmende die gezeigten Inhalte signifikant positiver als weibliche. Zugleich war mit steigendem Alter eine tendenziell strengere Beurteilung erkennbar.

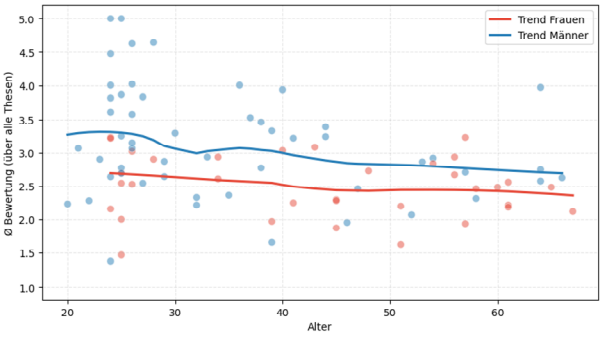

Abbildung 2 - Durchschnittliche Bewertung der Bildinhalte nach Alter und Geschlecht

Anschließend wurden auf der Plattform Midjourney verfügbare Bildbeispiele identifiziert, die auf Grundlage der geltenden Nutzungsbedingungen und Inhaltsrichtlinien als kritisch einzustufen sind (Midjourney o.J.). Die Einordnung der Befunde orientierte sich an risikobasierten Kategorisierungsansätzen für generative Modelle, wie sie in der „Topology of Risk“ beschrieben werden (Bird et al. 2023). Die qualitative Analyse ergab, dass problematische Inhalte trotz bestehender Richtlinien generiert und verbreitet werden, dass einzelne Inhalte gegen die Plattformvorgaben verstoßen und dass Nutzer*innen wiederkehrende Strategien zur Umgehung der Filter einsetzen, was auf substanzielle technische Lücken in der Durchsetzung der Richtlinien des Anbieters hinweist. Es zeigt sich, dass realitätsnahe synthetische Bilder mit geringem technischem Aufwand erzeugt werden können und ein erhebliches Potenzial für Desinformation, politische Einflussnahme und personenbezogene Rufschädigung bergen.

Die Zusammenführung beider Teilstudien zeigt, dass KI-Bilder durch Nutzer*innen nicht zuverlässig erkannt werden und KI-generierte sensible Motive von den Befragten überwiegend als ethisch problematisch bewertet werden. Zugleich sind vergleichbare Inhalte auf Plattformen verfügbar und teils trotz Verbot erzeugbar. Daraus ergibt sich konkreter Handlungsbedarf: Plattformrichtlinien müssen konsequent durchgesetzt werden, Filter sollten kontextsensitive Prüfungen unterstützen, Kennzeichnung und Herkunftsnachweise im Sinne des EU AI Acts sind verbindlich umzusetzen.

Literatur

Abdullah, Sami (2022), *Foto von Sami Abdullah auf Pexels*, Pexels https://www.pexels.com/de-de/foto/strasse-auto-vintage-mercedes-13818893/ (letzter Zugriff am 14.11.2025)

Bendel, Oliver (2025), *Image Synthesis from an Ethical Perspective,* AI & SOCIETY, 2025, Band 40, Auflage 2, https://doi.org/10.1007/s00146-023-01780-4

Bird, Charlotte, Ungless, Eddie L. und Kasirzadeh, Atoosa (2023) *Topology of Risks of Generative Text-to-Image Models*, in: AIES 23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, Montreal https://doi.org/10.48550/arXiv.2307.05543

Europäische Union (2024) *Verordnung über künstliche Intelligenz,* EUR-Lex https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689 (letzter Zugriff am 14.11.2025)

Hobelmann, Leon (2025) *Ethische Herausforderungen in der Bildgenerierung durch Künstliche Intelligenz am Beispiel von Midjourney,* Bachelorarbeit, Hochschule für Technik und Wirtschaft Berlin

Jobin, Anna, Ienca, Marcello und Vayena, Effy (2019) *The Global Landscape of AI Ethics Guidelines,* Nature Machine Intelligence, 2019, Band 1, Aufl. 9, S. 389–399 https://doi.org/10.1038/s42256-019-0088-2

Leow, Mikelle (2023) *Midjourney Bans Images Of China's Xi Jinping, Warns Users Not To Get Sneaky,* Designtaxi, URL: https://designtaxi.com/news/422929/Midjourney-Bans-Images-Of-Chinas-Xi-Jinping-Warns-Users-Not-To-GetSneaky/ (letzter Zugriff am 14.11.2025)

Midjourney (o.J.) *Community Guidelines,* Midjourney https://docs.midjourney.com/hc/en-us/articles/32013696484109-Community-Guidelines (letzter Zugriff am 14.11.2025)