LANDESANSTALT FÜR MEDIEN NRW
Der Meinungsfreiheit verpflichtet.

GENERATIVE KI FÜR ALLE?! ZWISCHEN EFFEKTIVITÄT UND VERANTWORTUNG

Ausgabe 16: Oktober 2024

fyi,

DER FORSCHUNGSMONITOR DER LANDESANSTALT FÜR MEDIEN NRW.
WISSENSCHAFTLICHE ERKENNTNISSE ZUM DIGITALEN WANDEL.

Realisiert von:
Dr. Valerie Hase & Philipp Knöpfle (M.Sc.),
Institut für Kommunikationswissenschaft und Medienforschung,
Ludwig-Maximilians-Universität München

GENERATIVE KI FÜR ALLE?! ZWISCHEN EFFEKTIVITÄT UND VERANTWORTUNG

Die rasante Entwicklung → **generativer künstlicher Intelligenz (KI)** markiert einen Wendepunkt für die Gesellschaft: →**Large Language Models (LLMs)**, wie ChatGPT oder Text-zu-Bild-Generatoren wie Stable Diffusion, MidJourney und DALL-E, revolutionieren wie wir Inhalte erstellen oder Wissen vermitteln. Diese KI-Technologien werden z. B. für Chat-Bots, zur Korrektur von Texten oder zur kreativen Erstellung eigener Multimedia-Inhalte genutzt. Dies birgt sowohl Chancen als auch Herausforderungen: Laut einer Umfrage aus dem Jahr 2024 bewerten 68 % aller Deutschen das Verhältnis von Nutzen und Risiken für KI als eher positiv. Jedoch fühlen sich nur 20 % mit der Technologie vertraut.[1] Fast die Hälfte der Internetnutzerinnen und -nutzer in Deutschland befürchtet laut einer weiteren Umfrage, dass die Menschheit die Kontrolle über KI verlieren könnte.[2] Doch was steckt eigentlich hinter dieser Technologie? Und sind solche Ängste gerechtfertigt?

Generative KI basiert auf komplexen Algorithmen, die in der Lage sind, große Datenmengen zu verarbeiten und daraus neue Inhalte zu erstellen. Durch den Einsatz maschinellen Lernens erlernen diese Modelle, Muster zu erkennen und eigenständig Texte, Bilder oder Videos zu generieren. Das bietet durchaus Chancen – z. B. für den Journalismus, der mithilfe von generativen KI-Prozesse automatisieren kann (siehe fyi 13).

Generative KI birgt jedoch auch Risiken: Insbesondere → **Deepfakes** – künstliche hergestellte Ton- oder Bilddateien, die z. B. reale Personen täuschend echt imitieren – stellen eine wachsende Gefahr dar. Sie werden zur Verbreitung von →**Desinformation** verwendet. Weitere Risiken zeigen sich im Bereich von →**Revenge Porn bzw. Deepfake Porn**, d. h. der Verbreitung von oft mit KI-generierten intimen Aufnahmen von Menschen ohne deren Zustimmung. Besorgniserregend ist, dass der deutschen Bevölkerung ethische Probleme im Hinblick auf KI bislang kaum präsent sind.[3]

AUF DEN PUNKT GEBRACHT

Generative Künstliche Intelligenz (KI): Computersysteme, die selbstständig neue Inhalte wie Texte, Bilder oder Musik erstellen, indem sie aus großen Mengen an vorhandenen Daten lernen.

Large Language Models (LLMs): KI-Systeme, die menschliche Sprache nachahmen bzw. erzeugen. Sie können z. B. Texte schreiben oder Fragen beantworten.

Deepfakes: gefälschte Bilder oder Videos, die mithilfe von künstlicher Intelligenz erstellt wurden, um Personen oder Ereignisse auf eine täuschend echte Weise dazustellen. Der Begriff „Deepfake" setzt sich aus den Worten „Deep Learning" und „Fake" zusammen.

Auch für demokratische Institutionen kann generative KI eine Gefahr bedeuten: Durch von KI generierter Desinformation könnten bspw. Wahlen beeinträchtigt werden – so zumindest die Befürchtung.[4] Zwar nimmt die EU mit dem AI Act eine Vorreiterrolle in der KI-Regulierung ein, doch bleiben Fragen zur praktischen Durchsetzbarkeit dieser Regulierung weiterhin offen.[5]

1 MeMo:KI (2022). Dashboard des Meinungsmonitor Künstliche Intelligenz. 01/2024, erstes Item „Risiko-Nutzen-Einschätzung: Gesellschaftlich", Wert von mind. 6 auf einer Skala von 0 (nur Risiko) bis 11 (nur Nutzen); 2. Item „Wissen: Sachverstand zu KI", Wert von mind. 4 auf einer Skala von 1 (Ich weiß nichts) zu 5 (Ich weiß sehr viel).

2 Schlude, A., Schwind, M., Mendel, U., Stürz, R. A., Harles, D., & Fischer, M. (2023). Verbreitung und Akzeptanz generativer KI in Deutschland und an deutschen Arbeitsplätzen.

3 Kieslich, K., Lünich, M., & Došenović, P. (2023). Ever Heard of Ethical AI? Investigating the Salience of Ethical AI Issues among the German Population.

4 Helberger, N., & de Vreese, C.H. (2024). Caught between Grand Ambitions and Small Print: The AI Act and the Use of GenAI in Elections und Łabuz., M., & Nehring, C. (2024). On the Way to Deep Fake Democracy? Deep Fakes in Election Campaigns in 2023.

5 Gstrein, O.J., Haleem, N., & Zwitter, A. (2024). General-Purpose AI Regulation and the European Union AI Act.

AUF DEN PUNKT GEBRACHT

Desinformation: bezeichnet die gezielte Verbreitung falscher oder irreführender Informationen, um die öffentliche Meinung zu manipulieren oder politische Ziele zu erreichen.

Revenge Porn bzw. Deepfake Porn: umfasst die absichtliche Verbreitung intimer Bilder oder Videos einer Person ohne deren Zustimmung, oft aus Rache oder mit dem Ziel, die betroffene Person zu demütigen oder zu schädigen. Entsprechende Inhalte werden oft mit KI-Technologien erstellt.

Die vierte Ausgabe des Forschungsmonitors 2024 (fyi 16) beleuchtet daher, wie generative KI die Gesellschaft herausfordert und welche Chancen bzw. Risiken mit dieser Technologie einhergehen.

Das sagt die Forschung: Generative KI ist bereits weit verbreitet – birgt aber Risiken

Studie 1: Generative KI – Fluch oder Segen für unsere Gesellschaft?
Generative KI beeinflusst unsere Gesellschaft auf einer Makro-Ebene (z. B. Demokratien durch Desinformation), auf einer Meso-Ebene (z. B. Prozesse in Unternehmen) und auf einer Mikro-Ebene (z. B. kreative Fähigkeiten).

Studie 2: Sechs Schlüsselfragen für unsere Zukunft mit generativer KI
Zentrale Herausforderungen der generativen KI umfassen Verzerrungen in solchen Modellen, die zu Diskriminierung führen können, unklare rechtliche Rahmenbedingungen sowie ökologische Bedenken.

Studie 3: Täuschung durch Technik? Generative KI und Fehlinformationen
Nutzerinnen und Nutzer verwenden oft einfache Faustregeln zur Beurteilung von KI-generierten (Fehl-)Informationen. Wenn KI-Modelle transparenter gestaltet werden, fällt es ihnen einfacher, Fehlinformationen zu erkennen.

Studie 4: Deepfakes entlarvt – Auswirkungen auf Medienvertrauen
Wenn Menschen erfahren, dass sie ein Deepfake gesehen haben, senkt dies ihr Vertrauen in die eigene Fähigkeit zur Erkennung solcher Fälschungen. Maßnahmen zur Medienkompetenz müssen kritisches Denken also fördern, ohne Bürgerinnen und Bürger grundsätzlich zu verunsichern.

Studie 5: Wie Jugendliche Risiken von KI erkennen und zu verstehen können
Die Medienkompetenz von Schülerinnen und Schüler im Hinblick auf generative KI kann durch Workshops gesteigert werden, in denen Jugendliche diese Technologien unter Anleitung ausprobieren und kritisch diskutieren.

Studie 6: Ein rechtlicher Flickenteppich? Zum Umgang mit Revenge Porn und Deepfake Porn in der EU
Rechtliche Grundlagen zur Verfolgung von „Revenge Porn“ und „Deepfake Porn“ variieren in den Mitgliedsstaaten der EU stark. Solche Inhalte werden auch und zunehmend durch generative KI hergestellt. Es braucht einen einheitlicheren Ansatz auf EU-Ebene, um den Schutz von Opfern zu verbessern.

INHALTSVERZEICHNIS

I. NEUE VERÖFFENTLICHUNGEN

STUDIE 1: GENERATIVE KI – FLUCH ODER SEGEN FÜR UNSERE GESELLSCHAFT?

Sætra, H. S. (2023). Generative AI: Here to Stay, But for Good? *Technology in Society*, *75*, 102372. https://doi.org/10.1016/j.techsoc.2023.102372

Open Access

Zentrale Fragestellung

Welche gesellschaftlichen Auswirkungen hat generative KI?

Methode

Literaturüberblick, d. h. keine empirische Methode.

Ergebnisse

Die Studie argumentiert, dass generative KI Chancen für menschliche Kreativität und Produktivität birgt. Es zeigen sich aber auch Herausforderungen: Auf gesamtgesellschaftlicher Makro-Ebene besteht die Gefahr, dass Demokratien durch Desinformation und gesellschaftliche Spaltung bedroht werden. Auf der Meso-Ebene einzelner Organisationen besteht das Risiko, dass Arbeitsplätze umgestaltet oder gestrichen werden. Auf individueller Mikro-Ebene kann der zunehmende Einsatz von KI zu einer Verringerung kreativer Fähigkeiten führen, weil Menschen diese weniger nutzen.

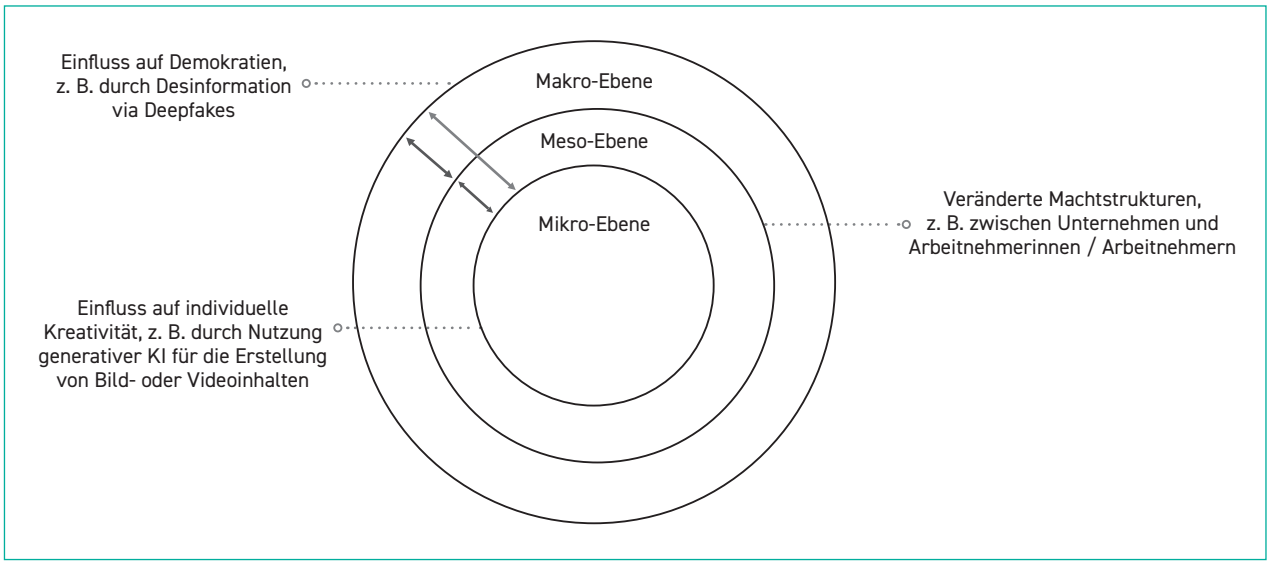

Abbildung 1. Gefahren generativer KI auf unterschiedlichen gesellschaftlichen Ebenen. Grafik von Sætra (2023).

Limitationen

Die Studie liefert einen Ausblick auf mögliche Chancen und Risiken, aber keine empirischen Daten dazu, wie Menschen tatsächlich auf generative KI reagieren. Ferner spricht der Artikel recht generell von „generativer KI", ohne auf unterschiedliche Formen dieser Technologie einzugehen.

Implikationen für die Praxis

Generative KI wird bereits jetzt in zahlreichen gesellschaftlichen Bereichen eingesetzt. Umso wichtiger ist es, mit dieser einhergehende Chancen und Risiken zu betrachten – und zwar auf Ebene der Gesellschaft, einzelner Organisationen und des Individuums, da sich mögliche Auswirkungen hier stark unterscheiden.

STUDIE 2: SECHS SCHLÜSSELFRAGEN FÜR UNSERE ZUKUNFT MIT GENERATIVER KI

Heaven, W. D. (2023). These Six Questions Will Dictate the Future of Generative AI. *MIT Technology Review.* https://www.technologyreview.com/2023/12/19/1084505/generative-ai-artificial-intelligence-bias-jobs-copyright-misinformation

Zentrale Fragestellung

Was sind die zentralen Herausforderungen generativer KI?

Methode

Literaturüberblick, d. h. keine empirische Methode.

Ergebnisse

Der Artikel beleuchtet sechs Herausforderungen mit Blick auf generative KI. Erstens beruhen diese Modelle auf Trainingsdaten, die oft Verzerrungen enthalten. Durch diese kann generative KI Diskriminierung fördern – z. B. generieren solche Modelle oft stereotype Beschreibungen von Männern („Ingenieure") und Frauen („Krankenschwestern"). Zweitens sind die rechtlichen Rahmenbedingungen für die Nutzung von urheberrechtlich geschütztem Material durch KI unklar. Drittens kann generative KI unterschiedliche Berufsfelder beeinflussen – obgleich sich hier durchaus Chancen zeigen, z. B. eine gesteigerte Produktivität durch die Nutzung von Chat-Bots. Insbesondere bei der KI-gestützten Erstellung von Inhalten besteht viertens die Gefahr der zunehmenden Verbreitung von Desinformation. Fünftens wirft die Nutzung generativer KI erhebliche ethische und ökologische Fragen auf, einschließlich des hohen Energieverbrauchs dieser Modelle. Sechstens, so der Autor, sind gesellschaftliche Diskussionen rund um KI häufig mit einer Art „Panikmache" verbunden, bei der primär Risiken dieser Technologie herausgestellt werden – teils aus Lobbyismus, etwa um mit den vermeintlichen existenziellen Konsequenzen von KI Profit zu machen oder um die Regulierung dieser Technologie zu beeinflussen.

Limitationen

Der Artikel skizziert Herausforderungen und Unsicherheiten im Umgang mit generativer KI. Dabei schlägt der Autor jedoch keine Lösungsansätze vor, um diese zu adressieren.

Implikationen für die Praxis

Die Nutzung generativer KI bringt Herausforderungen in Bereichen wie Politik und Wahlen, Medien und Urheberrecht oder der Berufswelt mit sich. Daher sind klare Richtlinien für den transparenten und ethischen Einsatz dieser Technologie dringend notwendig – zumal diese Technologie bereits jetzt genutzt wird.

STUDIE 3: TÄUSCHUNG DURCH TECHNIK? GENERATIVE KI UND FEHLINFORMATIONEN

Shin, D., Koerber, A., & Lim, J. S. (2024). Impact of Misinformation from Generative AI on User Information Processing: How People Understand Misinformation from Generative AI. *New Media & Society*. Online First Publication. https://doi.org/10.1177/14614448241234040

Zentrale Fragestellung

Wie nehmen Nutzerinnen und Nutzer Fehlformationen wahr, wenn sie generative KI wie ChatGPT nutzen?

Methode

Die Studie von Shin et al. untersucht anhand eines Experiments (N = 302), wie Nutzerinnen und Nutzer Fehlinformationen von generativen KI-Systemen verarbeiten. Dazu interagieren Teilnehmerinnen und Teilnehmer mit einer vermeintlichen KI-Plattform, die ChatGPT ähnelt. Die Forschenden testen dann, wie die Teilnehmenden auf (Fehl-) Information durch diese Plattform reagieren und ob erklärende Hinweise zur Plattform sich darauf auswirken, wie Teilnehmende mit dieser umgehen.

Ergebnisse

Nutzerinnen und Nutzer von generativer KI, etwa Chat-Bots, wenden oft einfache Faustregeln an, um die Glaubwürdigkeit von Informationen zu überprüfen. Wenn sie Informationen als für sich relevant und nützlich empfinden, sind sie eher in der Lage, die Genauigkeit von Informationen systematischer zu überprüfen. Erklärende Hinweise dazu, warum Plattformen wie ChatGPT bestimmte Informationen anzeigen, oder die Einbindung von Fakt-Checking sorgen zudem dafür, dass Nutzerinnen und Nutzer Informationen systematischer überprüfen. Entsprechend sollten diese Tools transparenter und nachvollziehbarer gestaltet werden. Gleichzeitig bedarf es einer entsprechenden Medienkompetenz von Nutzerinnen und Nutzern.

Limitationen

Die Studie wurde zu einem Zeitpunkt durchgeführt, als generative KI noch recht neu war. Wie viele weitere Studien könnten die Ergebnisse daher jetzt bereits überholt sein, da sich die entsprechende Technologie rasant weiterentwickelt. Zudem testet die Studie primär Informationssuche im Zusammenhang mit Fragen rund um die eigene Gesundheit – d. h. die Ergebnisse sind nicht für andere Formen von Informationssuche verallgemeinerbar.

Implikationen für die Praxis

Zur verantwortungsvollen Nutzung von generativer KI braucht es Interventionen im Hinblick auf Technologien, aber auch Nutzerinnen und Nutzer: Technologieunternehmen müssen ihre Systeme transparent und nachvollziehbar gestalten. Außerdem sollten Maßnahmen zur Verbesserung der Medienkompetenz von Nutzerinnen und Nutzer ergriffen werden, damit diese die entsprechenden Tools kritisch hinterfragen.

STUDIE 4: DEEPFAKES ENTLARVT – AUSWIRKUNGEN AUF MEDIENVERTRAUEN

Weikmann, T., Greber, H., & Nikolaou, A. (2024). After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media. *The International Journal of Press/Politics*. Online First Publication. https://doi.org/10.1177/19401612241233539

Open Access

Zentrale Fragestellung

Welchen Effekt hat es, wenn Menschen erfahren, dass sie von einem Deepfake getäuscht wurden?

Methode

Weikmann et al. zeigen Teilnehmenden in ihrer Experimentalstudie entweder einen echten oder einen mittels Deepfake gefälschten Nachrichtenbericht. Dabei wurde u. a. getestet, welchen Effekt verschiedene Formate (Audio, 2D-Video, 360°-Video) des Nachrichtenberichts und die Information darüber, ein Deepfake gesehen zu haben, auf das Vertrauen der Teilnehmenden haben.

Ergebnisse

Wenn Menschen erfahren, dass sie einen Deepfake gesehen haben, reduziert dies die Glaubwürdigkeit von Medienformaten deutlich. Auch das Vertrauen der Teilnehmenden in ihre Fähigkeit, Deepfakes zu erkennen, nahm dadurch geringfügig ab. Interessanterweise zeigen die Forschenden, dass das Medienformat keinen Einfluss auf die Glaubwürdigkeit der Deepfakes hatte. So wurden z. B. 360°-Videos nicht als glaubwürdiger wahrgenommen als „reguläre" 2D-Videos.

Limitationen

Der Artikel liefert spannende Ergebnisse im Hinblick auf Interventionen für einen möglichen Umgang mit Deepfakes. Allerdings könnte die kurze Zeit zwischen der Enttarnung des Deepfakes und nachfolgenden Fragen zum Vertrauen in die eigene Fähigkeit Deepfakes zu erkennen, die Ergebnisse verzerren. Zudem drehten sich die Nachrichtenbeiträge um das Thema Immigration – ein hoch polarisiertes Thema in Deutschland. Entsprechend könnten hier Voreinstellungen der Teilnehmenden eine Rolle spielen.

Implikationen für die Praxis

Die Studie zeigt, wie kompliziert Interventionen zur Steuerung von Medienkompetenz im Hinblick auf generative KI sind. Werden Bürgerinnen und Bürger damit konfrontiert, dass sie manipulierte Inhalte gesehen haben, kann dies z. B. ihr Vertrauen in die eigenen Fähigkeiten, diese zu erkennen, verringern. Maßnahmen zur Steigerung von Medienkompetenz müssen also Wege finden, eine kritische Grundeinstellung gegenüber digitalen Inhalten zu fördern – ohne Bürgerinnen und Bürger grundsätzlich zu verunsichern.

STUDIE 5: WIE JUGENDLICHE RISIKEN VON KI ERKENNEN UND VERSTEHEN KÖNNEN

Vartiainen, H., Kahila, J., Tedre, M., López-Pernas, S., & Pope, N. (2024). Enhancing Children's Understanding of Algorithmic Biases in and with Text-to-Image Generative AI. *New Media & Society*. Online First Publication. https://doi.org/10.1177/14614448241252820

Zentrale Fragestellung

Wie können Jugendliche für die Limitationen von KI sensibilisiert werden?

Methode

Vartiainen et al. führen drei Workshops mit Schülerinnen und Schülern (*N* = 209) in Finnland durch. In diesen Workshops wurden die Grundlagen von KI, die Entwicklung eigener KI-Anwendungen und die ethischen Implikationen von KI diskutiert und getestet. Dabei sollten die Jugendlichen ihre Eindrücke schriftlich festhalten. Zudem wurde getestet, wie gut die Schülerinnen und Schüler Risiken von KI nach den Workshops erkennen konnten.

Ergebnisse

Schülerinnen und Schüler waren durchaus in der Lage, Risiken in von KI erzeugten Bildern zu erkennen – etwa stereotype Verzerrungen im Hinblick darauf, wie unterschiedliche Geschlechter durch solche Programme dargestellt wurden. Oft war den Jugendlichen aber unklar, woher diese Verzerrungen stammen. Die interaktive Auseinandersetzung mit KI im Rahmen der Workshops verbesserte das Verständnis dieser Technologien und ihrer Verzerrungen dabei deutlich.

Limitationen

Die Ergebnisse sind nicht auf andere Altersgruppen oder Länder übertragbar, da in der Studie vergleichsweise junge Schülerinnen und Schüler aus Finnland untersucht wurden. Zudem zeigen sich Effekte vor allem für generative KI für die Erstellung von Bildern – unklar ist also, welche Folgen Medienkompetenz-Maßnahmen im Hinblick auf andere Technologien haben.

Implikationen für die Praxis

Effektive Bildungsstrategien zur Aufklärung über algorithmische Verzerrungen benötigen eine gezielte und altersgerechte Anleitung sowie praxisorientierte, kollaborative Lernaktivitäten. Wichtig ist, dass Jugendliche Chancen und Risiken dieser Technologien an konkreten Beispielen erproben und diskutieren. Solche Ansätze ermöglichen es Schülerinnen und Schülern, eine kritische Haltung gegenüber den ethischen und gesellschaftlichen Auswirkungen von KI zu entwickeln.

STUDIE 6: EIN RECHTLICHER FLICKENTEPPICH? ZUM UMGANG MIT REVENGE PORN UND DEEPFAKE PORN IN DER EU

Mania, K. (2024). Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study. *Trauma, Violence, & Abuse*, *25*(1), 117–129. https://doi.org/10.1177/15248380221143772

Zentrale Fragestellung
Wie werden Opfer von Revenge Porn und Deepfake Porn in der EU rechtlich geschützt?

Methode
Literaturüberblick, d.h. keine empirische Methode.

Ergebnisse
Die Studie diskutiert, wie Opfer von Revenge Porn (d. h. der Verbreitung intimer Aufnahmen ohne Zustimmung der Opfer) und Deepfake Porn (d. h. der Erstellung von pornographischen Materialien durch generative KI) in neun europäischen Ländern rechtlich behandelt werden. Insgesamt zeigt sich, dass der rechtliche Schutz mehr als unzureichend ist, vor allem im Vergleich mit einer deutlich pro-aktiveren Rechtsprechung in den USA: Während Länder wie Italien und die Niederlande spezifische Gesetze zur Strafverfolgung haben, betrachten andere Staaten wie Deutschland und Spanien solche Taten als Datenschutzverletzungen, was zu geringeren Strafen führt. In Deutschland gibt es z. B. keine explizite rechtliche Richtlinie zu Revenge Porn oder Deepfake Porn. Vielmehr werden existierende Gesetze, etwa die Datenschutzgrundverordnung, oder existierende Rechtsprechung zu einzelnen Fällen genutzt. Die Autorin fordert daher einen einheitlichen Ansatz auf EU-Ebene, um den Schutz der Opfer zu verbessern. Gleichzeitig bedarf es auch Maßnahmen durch digitale Plattformen, um die Verbreitung solcher Inhalte technisch zu verhindern.

Limitationen
Der Artikel bietet einen hilfreichen Überblick darüber, welche Regulierung in welchen EU-Ländern vorhanden ist – oder inwiefern diese fehlt. Hilfreich wäre hier ein konkreter Hinweis darauf, wie die laut Autorin vorbildliche Regulierung in den USA im Rahmen neuer Gesetzgebung, etwa des Digital Services Acts (DSA), in Europa umgesetzt werden könnte. Eine detailliertere Analyse, die Wege aufzeigt, wie der DSA oder ergänzende EU-weite Regelungen zur Bekämpfung von Revenge Porn beitragen könnten, würde die Studie weiter stärken.

Implikationen für die Praxis
In den Mitgliedsstaaten der EU sind die rechtlichen Regelungen zur Bekämpfung von Revenge Porn und Deepfake Porn lückenhaft und uneinheitlich. Sie schützen Opfer noch zu wenig. Während einige Länder bereits spezifische Gesetze z. B. zur strafrechtlichen Verfolgung haben, fehlt es an EU-weiten Regelungen. Zudem müssen auch Plattformen stärker gegen solche Inhalte vorgehen.

II. WAS SAGT DIE FORSCHUNG? INTERVIEW MIT DR. TERESA WEIKMANN

Deepfakes sind ein Risiko für die Gesellschaft – etwa für unser Vertrauen in die Medien oder Wahlentscheidungen

Interview mit Dr. Teresa Weikmann, Postdoktorandin an der Amsterdam School of Communication Research im BENEDMO-Projekt, welches die Verbreitung von Desinformation untersucht.

Was ist unter dem Begriff „Generative KI" zu verstehen?

Generative künstliche Intelligenz ist eine Art von KI, die in der Lage ist, neue oder originelle Inhalte zu erzeugen, anstatt nur auf vorhandene Daten zu reagieren oder diese zu analysieren. Vereinfacht gesagt, lernt die generative KI auf Basis existierender Muster und Strukturen, diese zu imitieren und auf neue Weise zusammenzusetzen. Dadurch können dann originelle Texte, Bilder, Musik, Videos oder sogar Programmcode entstehen.

Deepfakes sind in letzter Zeit immer häufiger in den Schlagzeilen. Wie genau funktionieren diese Technologien und was sind typische Beispiele für Deepfakes auf sozialen Medien?

Deepfakes fallen unter die Kategorie der generativen KI, weil sie basierend auf echten audio-visuellen Inhalten neue, gefälschte Inhalte erzeugen, die real aussehen oder klingen. Bei einem Deepfake-Video wird beispielsweise das Gesicht einer Person durch das Gesicht einer anderen Person ersetzt. Gleichzeitig kann ein Deepfake auch die Mimik und Sprache einer existierenden Person „lernen" und so realistisch nachahmen, dass dieser Person Worte in den Mund gelegt werden können. Typische Beispiele sind Videos von Politikern wie Olaf Scholz, in denen er scheinbar für ein AfD-Verbot plädiert. Allerdings tauchen Deepfakes vor allem auch im pornografischen Bereich auf, wodurch vor allem Frauen und andere vulnerable Gruppen betroffen sind.

Welche Herausforderungen gibt es bei der Erkennung von Deepfakes und wie effektiv sind Methoden zur Identifizierung dieser?

Es ist wichtig zu verstehen, dass sich Deepfakes in ihrer Qualität stark unterscheiden können. Zwar ist es technisch möglich, Deepfakes so realistisch zu gestalten, dass sie wie echte Aufnahmen wirken, doch dafür sind fortgeschrittene Programmierkenntnisse erforderlich. Frei verfügbare Programme zur Erstellung von Deepfakes liefern oft (noch) keine lebensnahen Ergebnisse, obwohl insbesondere im Audiobereich rasche Fortschritte gemacht werden. Schlecht gemachte Deepfakes lassen sich leicht erkennen, etwa durch unnatürliche Bewegungen im Videomaterial, wie Lippenbewegungen, die nicht mit der Tonspur übereinstimmen. Bei sehr gut gemachten Deepfakes, die mit dem bloßen Auge kaum von echten Videos zu unterscheiden sind, wird die Erkennung jedoch deutlich schwieriger. Es gibt zwar bereits Software zur Identifizierung von Deepfakes, aber diese ist oft noch nicht präzise genug. Wenn eine Erkennungssoftware beispielsweise angibt, dass ein Video zu 80 % echt ist, reicht das für Faktenprüferinnen und Faktenprüfer nicht aus. Sie benötigen 100 % Sicherheit.

Welche Effekte haben Deepfakes auf die Gesellschaft – z. B. im Hinblick auf Desinformation?

Deepfakes haben das Potenzial, Menschen so zu täuschen, dass sie falsche Inhalte für echt halten. Solche Täuschungen können politisches Verhalten beeinflussen, etwa bei einer Wahlentscheidung oder der Frage, wie stark ein Politiker oder eine Politikerin von der Öffentlichkeit unterstützt wird. Darüber hinaus kann das bloße Wissen um die Existenz von Deepfakes Unsicherheit verbreiten, wodurch Menschen insgesamt skeptischer gegenüber audiovisuellen

Nachrichteninhalten werden. Wenn alle Videos potenziell gefälscht sein könnten, kann dies das Vertrauen in visuelle Medien beeinträchtigen. Wer bereits einmal auf ein Deepfake hereingefallen ist, könnte zudem eine verstärkte Unsicherheit im Umgang mit Videos entwickeln.

Es gibt bereits einige Studien, die untersucht haben, ob Deepfakes im Vergleich zu textbasierter Fehlinformation oder authentischen Videos als glaubwürdig wahrgenommen werden. Die Ergebnisse zeigen, dass dies stark von der Qualität des Deepfake-Videos abhängt, wobei echte Videos nach wie vor eine höhere Glaubwürdigkeit genießen. Interessanterweise scheinen Deepfakes nicht unbedingt glaubwürdiger oder weniger glaubwürdig zu sein als textuelle Desinformation. Eine Studie, die ich mit meinen Kolleginnen an der Universität Wien durchgeführt habe, zeigt jedoch, dass die Irreführung durch Deepfakes dazu führt, dass Menschen audiovisuellen Medien insgesamt weniger vertrauen.

Welche Technologien existieren, um die Verbreitung von Deepfakes zu verhindern und wie könnten diese weiterentwickelt werden?

Als ich vor etwa zwei Jahren Interviews mit Faktenprüferinnen und Faktenprüfern zu diesem Thema geführt habe, haben sie mir erzählt, dass sie Deepfakes und andere Formen visueller Desinformation hauptsächlich mit manuellen Techniken untersuchen. Dabei gehen sie beispielsweise Frame für Frame durch ein Video und versuchen, Fehler oder Anzeichen von Manipulation zu erkennen. Seither hat sich einiges getan und die Technologien zur automatischen Erkennung haben sich stetig weiterentwickelt. Allerdings wird auch die Software zur Erstellung von Deepfakes immer besser – die beiden befinden sich in einem ständigen Wettrennen. Darüber hinaus wird auf EU-Ebene gerade darüber spekuliert, wie man die Verbreitung von Deepfakes am besten eindämmen kann, etwa durch die Kennzeichnung KI-generierter Inhalten mittels eines Wasserzeichens. Es bleibt jedoch unklar, wie effektiv diese Methoden tatsächlich sind.

III. FAZIT UND AUSBLICK

Fazit: Was wissen wir aus der aktuellen Forschung (noch) nicht?

Bei der Forschung zu generativer KI bestehen weiterhin wesentliche Forschungslücken. Erste Studien (siehe *Studie 1–2*) zeigen, dass generative KI bereits in ganz unterschiedlichen Teilen der Gesellschaft angewendet wird – mit entsprechenden Chancen und Risiken. Allerdings wurden mögliche Auswirkungen in verschiedenen kulturellen Kontexten bisher wenig untersucht. Auch ist unklar, welche Effekte unterschiedliche Formen generativer KI (z. B. Chat-Bots, Tools zur Generierung von Bildern) haben. Ferner weisen erste Analysen (siehe *Studie 3–4*) darauf hin, dass die Nutzung generativer KI z. B. für Deepfakes und die Diskussion darüber dazu führen könnte, dass Menschen digitalen Informationen und ihren eigenen Fähigkeiten zur Erkennung von Fehlinformationen weniger vertrauen. Unklar bleibt jedoch, wie langfristig diese Effekte sind. Lösungsansätze für einen kritischen, aber konstruktiven Umgang mit generativer KI umfassen Medienkompetenz-Maßnahmen, das Monitoring entsprechender Inhalte durch Plattformen, Transparenzhinweise in KI-Tools und eine aktualisierte Gesetzgebung (siehe *Studie 3, 5–6*).

Ausblick

Auf Basis dieser Forschungsdesiderate – welche Aspekte könnte die Forschung, aber auch die Medienpraxis zukünftig adressieren?

Ausblick 1: Stärkung der Medienkompetenz für den Umgang mit KI

Die Implementierung von Bildungsprogrammen zu KI ist entscheidend. Durch gezielte Bildungsinitiativen können Bürgerinnen und Bürger, insbesondere junge Menschen, ein fundiertes Verständnis für die Funktionsweise von und Medienkompetenz für den Umgang mit KI entwickeln. Solche Programme sollten nicht nur technisches Wissen z. B. über Verzerrungen in diesen Modellen vermitteln, sondern insbesondere das kritische Denken und die Fähigkeit fördern, zwischen echten und manipulierten Informationen zu unterscheiden.

Ausblick 2: Transparenz und Regulierung für KI-Tools

Auf technischer Ebene sollten KI-Tools transparenter gestaltet werden, damit z. B. die Authentizität von Informationen klar erkennbar ist und die Herkunft von Informationen transparent gemacht wird. Entsprechende wissenschaftliche Debatten gibt es z. B. im Hinblick auf sogenannte erklärbare künstliche Intelligenz. Um die Transparenz von Tools zu standardisieren und ein Monitoring durchzuführen, inwiefern dies eingehalten wird, bedarf es jedoch einer länderübergreifenden Regulierung. Hier müssen Forschende, Plattformen und Regulierungsunternehmen stärker zusammenarbeiten, um mit der sich stetig veränderten Technologie Schritt zu halten.

IV. FORSCHUNGSPROJEKTE

Künstliche Intelligenz als vertrauenswürdige:r Journalist:in? Effekte KI-generierter Botschaften auf skeptische Bürger:innen

Zentrale Fragestellung: **Wie kann KI dazu beitragen, die Annahmebereitschaft von Klima-Botschaften zu erhöhen?**
Das Projekt untersucht, wie Bürgerinnen und Bürger KI-generierte journalistische Inhalte zum Klimawandel wahrnehmen. Durch Experimente, Befragungen und Inhaltsanalysen wird geprüft, ob KI als eine Art neutrale Vermittlerin den Dialog und die Offenheit in polarisierten Debatten fördern kann oder hinderlich ist.

Projektteam: Universität Passau (Prof. Hannah Schmid-Petri), Bayrisches Forschungsinstitut für digitale Transformation (Daria Kravets-Meinke).

Generative Künstliche Intelligenz in der Arbeitswelt (GENKIA)

Zentrale Fragestellung: **Wie beeinflusst generative KI die Arbeitswelt in Deutschland?**
Das Projekt untersucht die Auswirkungen generativer KI auf fünf zentrale Berufsfelder: Marketing, Human Resource Management, Programmierung, Journalismus und öffentliche Verwaltung. Ziel ist es, die Erfahrungen von Beschäftigten mit generativer KI zu erfassen, deren Einfluss auf die Beschäftigung und die Arbeitsqualität zu analysieren und Handlungsempfehlungen zu entwickeln.

Projektteam: Wissenschaftszentrum Berlin für Sozialforschung (Prof. Florian Butollo, Dr. Christine Gerber, Ann-Kathrin Katzinski, Marlene Cäcilie Kulla, Mareike Sirman-Winkler) in Kooperation mit dem Alexander von Humboldt Institut für Internet und Gesellschaft.

Impressum

Herausgeberin:
Landesanstalt für Medien NRW
Zollhof 2
40221 Düsseldorf
Tel: 0211 / 77 00 7- 0
Fax: 0211 / 72 71 70
E-Mail: info@medienanstalt-nrw.de
Direktor: Dr. Tobias Schmid

Projektleitung:
Dr. Meike Isenberg
(Leitung Medienpolitik und Forschung)
Sabrina Nennstiel
(Leitung Kommunikation)

Realisierung:
Dr. Valerie Hase
Institut für Kommunikationswissenschaft
und Medienforschung (IfKW)
Ludwig-Maximilians-Universität München
Akademiestr. 7
80799 München

Projektleitung:
Dr. Valerie Hase
(IfKW LMU München)

Autor und Redaktion:
Philipp Knöpfle (M.A.)

Gestaltung:
Merten Durth (DISEGNO kommunikation)

Herausgegeben:
Oktober 2024

„FYI – der Forschungsmonitor der Landesanstalt für Medien NRW" wird durch das Team von Dr. Valerie Hase vom Institut für Kommunikationswissenschaft und Medienforschung der Ludwig-Maximilians-Universität München erarbeitet.

Ziel dieses Forschungsmonitors ist es, aktuelle Entwicklungen im Themenfeld digitale Informationsintermediäre und öffentliche Meinungsbildung so aufzubereiten, dass das Monitoring einen Überblick über neue wissenschaftliche Publikationen, den Stand aktueller Forschungsprojekte und kommende relevante Veranstaltungen im Themenfeld verschafft.

Philipp Knöpfle und Dr. Valerie Hase sind wissenschaftliche Mitarbeiter bzw. Mitarbeiterinnen am Institut für Kommunikationswissenschaft und Medienforschung der LMU München. Sie arbeiten hier u. a. an Forschungsprojekten zu Forschungsethik, Open Science, digitalem Journalismus sowie automatisierten Methoden.