ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

PROJEKTBERICHT

WIEN, JÄNNER/2025
ITA-2025-01
WWW.OEAW.AC.AT/ITA

GENERATIVE KI UND DEMOKRATIE

ENDBERICHT JÄNNER 2025

itas

ITA

GENERATIVE KI UND DEMOKRATIE

ENDBERICHT JÄNNER 2025

Institut für Technikfolgen-Abschätzung
der Österreichischen Akademie der Wissenschaften
Projektleitung: Michael Nentwich
Autor:innen: Michael Nentwich
Steffen Bettin
Saskia Favreuille
Fabian Fischer
Jutta Jahnel (ITAS)
Jaro Krieger-Lamina
Walter Peissl

Studie im Auftrag des österreichischen Parlaments

Wien, Jänner/2025

IMPRESSUM

Medieninhaber:
Österreichische Akademie der Wissenschaften
Juristische Person öffentlichen Rechts (BGBl 569/1921 idF BGBl I 31/2018)
Dr. Ignaz Seipel-Platz 2, A-1010 Wien

Herausgeber:
Institut für Technikfolgen-Abschätzung (ITA)
Bäckerstraße 13, A-1010 Wien
www.oeaw.ac.at/ita

Die ITA-Projektberichte erscheinen unregelmäßig und dienen der Veröffentlichung der Forschungsergebnisse des Instituts für Technikfolgen-Abschätzung.
Die Berichte erscheinen in geringer Auflage im Druck und werden über das Internetportal „epub.oeaw" der Öffentlichkeit zur Verfügung gestellt:
epub.oeaw.ac.at/ita/ita-projektberichte

ITA-Projektbericht Nr.: ITA-2025-01 (Wien, Jänner/2025)
ISSN: 1819-1320
ISSN-online: 1818-6556
epub.oeaw.ac.at/ita/ita-projektberichte/ITA-2025-01.pdf

INHALT

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

VERZEICHNIS DER BOXEN (KÄSTEN)

KURZFASSUNG

Dieser Bericht stellt die Ergebnisse einer Technikfolgenabschätzungsstudie im Auftrag des österreichischen Parlaments dar. Ausgangspunkt war die Beobachtung, dass sich die relativ neue Technologie Generative Künstliche Intelligenz (KI) seit 2022 in großer Geschwindigkeit in vielfältige Lebensbereiche verbreitet. In manchen Bereichen wird ihr disruptives Potenzial attestiert und auch für die politische Sphäre werden nicht nur Chancen, sondern auch große Risiken gesehen. Dementsprechend stellt der vorliegende Bericht diese Chancen Generativer KI für die Demokratie im Überblick dar und widmet sich im Hauptteil der Analyse der vielfältigen Risiken für die politische Meinungsbildung und den öffentlichen Diskurs. Der Bericht beginnt einleitend mit einer Darstellung des Stands der Technik, insbesondere von Text-, Bild-, Video- und Tongeneratoren und technischen Ansätzen, um KI-generierte Inhalte zu erkennen. Im Einleitungskapitel werden auch grundlegende, untersuchungsrelevante Begriffe wie Desinformation, Fake News und Co. erläutert. Ein eigenes Kapitel ist weiteren gesellschaftsrelevanten Folgen Generativer KI, etwa in den Bereichen Umwelt und Arbeit, gewidmet. Im letzten Teil werden Handlungsoptionen in den Bereichen Regulierung, Organisation und Technik erörtert und schließlich Schlussfolgerungen für das österreichische Parlament und die interessierte Öffentlichkeit gezogen. In den folgenden Absätzen werden alle Kapitel für die eilige Leserschaft kurzgefasst.

Bericht analysiert Chancen und Risiken Generativer KI für die Demokratie …

… erörtert Handlungsoptionen …

… und zieht Schlussfolgerungen für das Parlament

Das Vorwort beschreibt den Bezugspunkt für die nachfolgende Analyse, nämlich die liberale Demokratie österreichischer Prägung, in der dem freien öffentlichen Diskurs als Grundlage einer unbeeinflussten Willensbildung eine zentrale Rolle zukommt. Dieser wird auch durch die zunehmende Digitalisierung herausgefordert, etwa durch die teilweise Verlagerung in soziale Medien. Die Verbreitung falscher oder irreführender KI-generierter Informationen zeichnet sich als neues Problemfeld ab. Im Einleitungskapitel werden daher verschiedene Begriffe rund um das Phänomen der Diskursverzerrung erörtert, insbesondere Fake News, Deepfakes, Desinformation und Misinformation u.v.m. Hauptgegenstand der vorliegenden Studie ist das Phänomen der Desinformation, also der intendierten (absichtlichen) Verbreitung unwahrer oder irreführender Informationen.

[Vorwort] Analyserahmen Demokratie

[1.2] Wichtige Begriffe im Zusammenhang mit Diskursverzerrungen

Angesichts der enormen Dynamik der Technikentwicklung in diesem Feld ist es schwierig, überhaupt von einem „Stand“ der Technik zu sprechen, ist doch fast jede solche Beschreibung wenig später bereits überholt. Dennoch geben die entsprechenden Berichtsabschnitte einen guten Einblick in die sich stets ausweitenden Möglichkeiten Generativer KI. Im Zusammenhang mit dem Thema dieser Studie ist es wichtig zu wissen, dass es mittlerweile praktisch ohne Vorkenntnisse und mit geringem Mitteleinsatz möglich ist, diese Technologien zu nutzen, um zumindest Nicht-Expert:innen zu täuschen. Demgegenüber dürfte es auch langfristig kaum möglich bleiben, KI-generierte Inhalte auf technischem Weg eindeutig als solche zu identifizieren. Den Potenzialen stehen aus technischer Sicht auch schwerwiegende Begrenzungen gegenüber: So sind sog. Halluzinationen, also Erfindungen durch Generative KI, die nicht faktenbasiert sind, prinzipiell nicht auszuschließen und dementsprechend zu berücksichtigen.

[2] Dynamischer „Stand“ der Technik

Halluzinationen technisch nicht zu vermeiden

Die Untersuchung der Potenziale Generativer KI für die Demokratie hat interessante Anwendungsmöglichkeiten in vier Bereichen ergeben. Erstens gibt es Einsatzmöglichkeiten in den (klassischen) Medien, bspw. als Recherchetool und

[3] Chancen für die Demokratie

zur Erstellung personalisierter News. Zweitens kann Generative KI zur Demokratisierung von Information dienen, etwa zur Sprachübersetzung und -vereinfachung und für automatisierte Zusammenfassungen. Drittens gibt es ein Potenzial zur Diskursverbesserung, etwa bei KI-unterstützter Strukturierung von Diskussionen oder bei der vereinfachten Kontaktaufnahme mit Politiker:innen. Viertens konnten auch etliche potenzielle Werkzeuge für den politischen Alltag gefunden werden, z. B. zur automatisierten Protokollerstellung oder als Recherchetool.

[4.1] Verschärfung bestehender Diskursverzerrungen …

Das Hauptkapitel des Berichts analysiert zunächst die spezifischen Risiken Generativer KI für den politischen Diskurs. Dabei wurde herausgearbeitet, dass Generative KI einerseits das Potenzial hat, bereits bestehende Verzerrungen zu verschärfen (Hassrede, Echokammern usw.), nämlich durch die erleichterte Erstellung irreführender, schädigender und polarisierender Inhalte und von Fake Accounts für Desinformationskampagnen. Darüber hinaus gibt es aber auch zusätzliche spezifische Folgen der Generativen KI: Ausgehend von den zum Training verwendeten Daten kann es einerseits zu systematischen Verzerrungen (Bias) kommen, andererseits kann auch Misinformation „gelernt" und in Folge weiterverbreitet werden. Mittelfristig ist eine Einschränkung der Diversität im Diskurs zu erwarten. Dazu kommt die prinzipielle mangelnde Verlässlichkeit der Textgeneratoren und damit die Gefahr der Misinformation durch sog. Halluzinationen.

… und qualitativ neue Risiken durch Generative KI

[4.2] Disruptives Potenzial von Deepfakes in der Politik

Deepfakes (täuschend echt wirkende Bilder bzw. Video- und Audioaufnahmen, die aber künstlich erstellt sind) haben darüber hinaus disruptives Potenzial, da insbesondere Politiker:innen eine potenzielle Zielgruppe sind. Die Diskreditierung von Personen führt zu psychologischen und finanziellen Folgen, jene von Parteien, Journalismus und Medien zu Glaubwürdigkeitsverlust und der einhergehenden Schwächung demokratischer Institutionen. Schließlich können Deepfakes ein wirksames Mittel für Propaganda sein. Zu all diesen Potenzialen gibt es bereits etliche (internationale) Beispiele, die im Bericht auch beispielhaft dokumentiert werden. Es kann jedoch bis dato in Europa bzw. Österreich noch nicht davon gesprochen werden, dass diese Tools bereits im großen Stil zum Einsatz kämen und schon jetzt ein massives Problem darstellten.

[4.3] Microtargeting

Politisches Microtargeting, also das Ansprechen fein ausdifferenzierter Zielgruppen, wird nicht nur als Chance, sondern auch als Risiko für die demokratische Auseinandersetzung diskutiert. Generative KI hat das Potential, die bereits existierende Praxis deutlich effizienter zu machen, indem die psychologische Zielgenauigkeit durch die Überzeugungskraft von (generiertem) Bildmaterial noch deutlich erhöht wird und die Messages und deren Verbreitungswege völlig individualisiert werden können. Microtargeting kann aber auch zur bewussten Manipulation eingesetzt werden kann, etwa indem widersprüchliche zielgruppenspezifische Inhalte versendet werden. Die meisten Beobachter:innen und sogar der europäische Gesetzgeber kommen zum Schluss, dass die Effekte von Microtargeting aus ethischer und demokratiepolitischer Sicht problematisch sind, das EU-Gesetz über Transparenz und Targeting politischer Werbung ist ein erster Schritt der Regulierung dieser Materie.

[4.4] Kommerzielle Machtkonzentration im Sektor Generative KI

Zu den Auswirkungen von Generativer KI zählt auch die zunehmende wirtschaftliche Machtkonzentration im Bereich der KI-Entwicklung und des Angebots an entsprechenden Anwendungen. Eine Konzentration im digitalen Markt ist schon seit langem beobachtbar und auf Netzwerk- und Skaleneffekte zurückzuführen. Im neuen Bereich der Generativen KI sind sowohl bekannte als auch

neue Firmen aktiv. Insbesondere Infrastrukturunternehmen wie Chiphersteller und Rechenzentren gewinnen aufgrund der hohen Rechenleistung an Bedeutung. Techno-ökonomische Ursachen wie der Bedarf hoher Rechenleistung und großer Datenmengen zum Training Generativer KI verstärken weiters die Tendenz zu Machtkonzentration bei Digitalunternehmen. Mittlerweile dominieren IKT-Unternehmen sogar staatliche Akteure, was auch im militärischen Bereich eine zunehmende Rolle spielt. Somit entstehen neue Oligopole, die Digitalkonzerne vergrößern ihre Macht und dringen in neue gesellschaftliche Sphären vor.

[4.5] Hybride Bedrohungen

Es sind aber nicht nur kommerzielle Akteure, die das globale Kräftespiel aufgrund der Einsatzmöglichkeiten von Generativer KI verändern, sondern auch staatliche sowie kriminelle Akteure. Unter hybriden Bedrohungen werden Aktivitäten ausländischer Akteure verstanden, die nicht direkt auf klassische militärische Maßnahmen setzen, jedoch im Krieg oder zu dessen Vorbereitung sowie der gegnerischen Abschreckung eingesetzt werden. Ziel ist es beispielsweise, die öffentliche Meinung in der gegnerischen Bevölkerung durch Desinformation zu beeinflussen und damit die öffentliche Ordnung zu destabilisieren, Kommunikationsinfrastrukturen anzugreifen und das Vertrauen in demokratische Strukturen zu schwächen. Generative KI ist ein einfach zu skalierendes, effektives Mittel, dies zu erreichen, ohne überhaupt ins angegriffene Land vordringen zu müssen.

[4.6] Cyberkriminalität

Aber auch Cyberkriminalität in unterschiedlichen Formen, etwa durch Erpressung und Einschüchterung politischer Akteure spielt eine zunehmende Rolle. Kriminelle nutzen Generative KI zur Imitation und zur Arbeitserleichterung, etwa für gezielte Täuschung. Es sind aber auch Angriffe auf die Generative KI selbst in verschiedenen Varianten vorstellbar, etwa indem die Daten, auf die die KI zugreift, abgesaugt werden oder verfälscht werden.

[4.7] Digitale Souveränität bedroht

Gemeinsam stellen diese aufgelisteten Risiken eine multiple Bedrohung der digitalen Souveränität demokratischer Systeme aber auch der einzelnen Bürger:innen dar: Der Missbrauch Generativer KI kann Wahlen und Meinungsbildung manipulieren, das Vertrauen in den Journalismus schwächen, die Gesellschaft weiter polarisieren und Bürger:innen über politische Entscheidungsprozesse in die Irre führen. Zugleich verkleinert die Machtkonzentration der Anbieter von Generativer KI den Spielraum der Staaten und es können hybride Bedrohungen von außerhalb des Staates und innere Bedrohungen durch Cyberkriminalität zunehmen.

[5] „Sonstige" Folgen nicht weniger beachtlich

Diese Studie fokussiert auftragsgemäß auf die Folgen im politischen Bereich im engeren Sinne, daher wurde über nicht auf die Demokratie bezogenen Auswirkungen nur ein grober Überblick erarbeitet. Damit soll der irreführende Eindruck vermieden werden, dass Generative KI nur in Bezug auf Demokratie Chancen und Risiken birgt. Sind doch die „sonstigen" Folgen zum Teil so gravierend, dass ein bloßer Fokus auf das politische System zu kurz greifen würde. Dies betrifft die Bereiche Umwelt, Arbeitswelt, Bildung, Gesellschaft und Recht. Hier sollen nur beispielhaft der enorme Energiehunger und die damit verbundenen CO_2-Emissionen der Rechenzentren für KI sowie deren Kühlwasser- und Flächenverbrauch und Fragen des Urheberrechts hervorgehoben werden.

[6] Zusammenstellung und Analyse der Handlungsoptionen:
– regulativ
– organisatorisch
– technisch

Vor dem Hintergrund der Analyse der Chancen und Risiken Generativer KI in Hinblick auf das demokratische politische System wurde in der Studie abschließend die Frage gestellt, ob die Gesellschaft die angestoßene sozio-technische Entwicklung passiv beobachten und gewähren lassen könne oder ob es vielmehr angezeigt wäre, den Prozess zu gestalten, d. h. in der einen oder anderen Weise versuchen einzugreifen. Zur Beantwortung dieser Frage wurde in einer

breiten Recherche nach Handlungsoptionen gesucht, die darauf hinauslaufen, erwünschte Entwicklungsrichtungen zu fördern und erkannte Risiken auszuschließen oder zumindest zu minimieren. In dieser Studie wird zwischen regulativen, organisatorischen und technischen Ansätzen unterschieden.

[7] Schlussfolgerungen für das Parlament:

Diese zahlreichen, in der wissenschaftlichen und populärwissenschaftlichen Literatur, in Policypapers, aber auch in Rechtsakten und den Medien gefundenen Optionen wurden in einem mehrstufigen Verfahren strukturiert und analysiert und schließlich für den primären Adressaten österreichisches Parlament verdichtet. Auf diese Weise wurden folgende wesentliche Schlussfolgerungen herausgearbeitet:

Demokratischen Diskurs stärken

- Parlamentarische Enquetekommission „Demokratie und KI" einrichten
- Verhaltenskodex zu KI in der Politik erarbeiten
- Bundesweit Bürger:innen-Foren zu Grundsatzfragen der Demokratie abhalten
- Medien- und KI-Literacy massiv fördern
- Transparenz durch Kennzeichnung und Förderung der Ansätze zu Erklärbarkeit erhöhen
- Spezifische Regulierungsvorschläge diskutieren, insb.:
  - Verantwortlichkeit der Plattformen für Inhalte festlegen
  - Politisches Microtargeting generell verbieten
  - Politisch motivierte Deepfakes mit Schadensabsicht verbieten

Digitale Souveränität auf demokratischen Prinzipien aufbauen

- Österreich in der EU als Vorreiter einer proaktiven Vorgangsweise zum Erhalt der Demokratie positionieren
- Initiativen in Richtung staatliche digitale Souveränität setzen
- Demokratieverträgliche europäische Diskursplattform aufbauen

- Manipulationsversuche systematisch und konsequent abwehren, insb.:
  - Fact-Checking fördern
  - Koordination gegen feindliche Einmischung intensivieren
  - Umfassende Kennzeichnung durchsetzen

Chancen ausloten, Begleitforschung fördern

- Entwicklung chancenreicher KI-Anwendungen im politischen Kontext fördern
- KI-Begleitforschung forcieren
- Jährlichen Monitoringbericht zur digitalen politischen Kommunikation in Österreich erstellen

Optionen-Mix notwendig

Diese Fülle an Optionen begründet sich auch aus der Einschätzung, dass Regulierung alleine nicht ausreicht, sondern eine Kombination unterschiedlicher, gut aufeinander abgestimmter Maßnahmen (regulative, organisatorische und technische) notwendig ist, um den durch Generative KI beförderten Risiken für die Demokratie wirksam begegnen zu können.

Vorschlag Leseanleitung für Eilige

Aufgrund des komplexen Themas kann der vorliegende Bericht mit seinen rund 125 Seiten als eine Art Referenzrahmen gesehen werden kann, der es ermöglicht, bestimmte Details nachzulesen. Für die eilige Leserin, den eiligen Leser schlagen die Autor:innen vor, noch die thematische Einführung (Abschnitt 1.1 – 4 Seiten) sowie die Schlussfolgerungen (Kapitel 7 – 6 Seiten) genauer zu lesen und den restlichen Text anhand der Marginalien zu überfliegen.

EXECUTIVE SUMMARY

This report presents the findings of a technology assessment study commissioned by the Austrian Parliament addressing the effects of Generative Artificial Intelligence (AI) on democracy. This relatively new technology gained momentum in 2022 and has spread rapidly into many areas of life. In some areas, its disruptive potential has been recognised, and not only opportunities but also major risks are seen in the political sphere. Accordingly, this report provides an overview of the opportunities of Generative AI for democracy and dedicates the main section to analysing the various risks for political opinion-forming and public discourse. The report begins by describing the state of the art, mainly of text, image, video, and sound generators, as well as technical approaches to recognising AI-generated content. The introductory chapter also explains basic terms relevant to the study, such as disinformation, fake news, etc. A separate chapter is dedicated to other socially relevant consequences of Generative AI, such as those in the environment and labour. The final section discusses options for action in regulation, organisation, and technology and concludes with a particular focus on the Austrian Parliament and the interested public. In the following paragraphs, all chapters are summarised for readers in a hurry.

The report analyses the opportunities and risks of Generative AI for democracy …

… discusses options for action …

… and draws conclusions for Parliament

The foreword describes the point of reference for the following analysis, namely liberal Austrian-style democracy, in which free public discourse plays a central role as the basis for uninfluenced decision-making. This is also challenged by increasing digitalisation, for example, through the partial shift to social media. The dissemination of false or misleading AI-generated information is emerging as a new problem area. The introductory chapter, therefore, discusses various terms relating to the phenomenon of discourse distortion, in particular fake news, deepfakes, disinformation and misinformation, and much more. The main subject of this study is the phenomenon of disinformation, i.e., the intentional dissemination of untrue or misleading information.

[Foreword] Framework for analysing democracy

[1.2] Essential terms in connection with discourse bias

Given the enormous dynamics of technological development in this field, it is difficult to speak of a „state of the art“ at all, as almost every such description is already outdated a short time later. Nevertheless, the relevant sections of the report provide an insight into the ever-expanding possibilities of Generative AI. In the context of the topic of this study, it is essential to know that it is now possible to use these technologies to deceive at least non-experts with virtually no prior knowledge and with little investment of resources. In contrast, it might soon be impossible to identify AI-generated content as such technically. From a technical perspective, there are also severe limitations to the potential: Hallucinations, i.e., inventions by Generative AI that are not fact-based, cannot be ruled out in principle and must be taken into account accordingly.

[2] Dynamic „state“ of the art

Hallucinations technically unavoidable

The investigation into the potential of Generative AI for democracy has revealed interesting applications in four areas. First, there are opportunities in the (traditional) media, for example, as tools for research tool and creating personalised news. Second, Generative AI can democratise information through language translation, text simplification, and automated summaries. Third, there is potential for improving discourse, such as AI-supported structuring of discussions or simplified contact with politicians. Finally, several tools can make political life more manageable, e.g., the automated creation of minutes or as a research tool.

[3] Opportunities for democracy

[4.1] Exacerbation of existing discourse distortions …

The central chapter of the report starts by analysing the specific risks of Generative AI for political discourse. It has become apparent that Generative AI has the potential to exacerbate existing distortions (hate speech, echo chambers, etc.) by facilitating the creation of misleading, harmful and polarising content and fake accounts for disinformation campaigns. In addition, there are also specific consequences of Generative AI: on the one hand, the data used for training can lead to systematic distortions (bias), and on the other hand, misinformation can also be „learnt" and subsequently disseminated. In the medium term, a restriction of diversity in the discourse is to be expected. Added to this is the fundamental lack of reliability of the text generators, heightening the risk of misinformation through so-called hallucinations.

… and qualitatively new risks through Generative AI

[4.2] Disruptive potential of deepfakes in politics

Deepfakes – deceptively real-looking images or video and audio recordings that are artificially created – also have disruptive potential, as politicians, mainly, are a potential target group. Discrediting individuals leads to psychological and financial consequences, while discrediting political parties, journalism, and the media leads to a loss of credibility and the associated weakening of democratic institutions. Finally, deepfakes can be an effective means of propaganda. Several (international) examples of all these potentials are already documented in the report. However, it is not yet possible to say that these tools are already being used on a large scale in Europe or Austria and already represent a massive problem.

[4.3] Microtargeting

Political microtargeting, i.e., addressing finely differentiated target groups, is being discussed as an opportunity and a risk for democratic debate. Generative AI has the potential to make existing practices much more efficient by significantly increasing the psychological accuracy of targeting through the persuasive power of (generated) image material and completely individualising the messages and their distribution channels. However, microtargeting can also be used for deliberate manipulation, such as sending contradictory target group-specific content. Most observers and even European legislators have concluded that the effects of microtargeting are problematic from an ethical and democratic point of view, and the EU law on transparency and targeting of political advertising is a first step towards regulating this issue.

[4.4] Commercial concentration of power in the sector Generative AI

The effects of Generative AI also include the increasing concentration of economic power in AI development and the range of corresponding applications. In the digital market, such concentration has been observable for a long time and can be attributed to networks and economies of scale. Established and new companies are active in the new field of Generative AI. Infrastructure companies, such as chip manufacturers and data centres, are gaining importance due to their high computing power. Techno-economic reasons such as the need for high computing power and large amounts of data to train Generative AI also reinforces the trend towards a concentration of power among digital companies. ICT companies now dominate state actors and play an increasing role in the military sector. As a result, new oligopolies are emerging, and digital companies are increasing their power and gaining access to new spheres of society.

[4.5] Hybrid threats

However, it is not only due to commercial actors using Generative AI that the global landscape is changing. The use of AI by state and criminal actors plays also a role. Hybrid threats are activities by foreign actors that do not rely directly on traditional military measures but are used in war or preparation for war and to deter the enemy. The aim is, for example, to influence public opinion in the opposing population through disinformation and thus destabilise public order,

attack communication infrastructures and weaken trust in democratic structures. Generative AI is an easily scalable, effective means of achieving this without even having to enter the country under attack. However, cybercrime in various forms, such as blackmail and intimidation of political actors, is also playing an increasingly important role. Criminals use Generative AI to imitate and facilitate work, for example, for targeted deception. However, attacks on Generative AI are also conceivable in various forms, such as siphoning off or falsifying the data the AI accesses.

[4.6] Cybercrime

Together, all these risks pose multiple threats to the digital sovereignty of democratic systems and individual citizens: The misuse of Generative AI can manipulate elections and opinion-forming, weaken trust in journalism, further polarise society and mislead citizens about political decision-making processes. At the same time, the concentration of power of Generative AI providers reduces the room for manoeuvre of states and hybrid threats from outside the state, and internal threats from cybercrime may increase.

[4.7] Digital sovereignty under

In line with its remit, this study focuses on the consequences in the political sphere in a narrower sense, which is why only a rough overview of effects unrelated to democracy has been compiled. This is intended to avoid the misleading impression that Generative AI only harbours opportunities and risks to democracy. However, in some cases, the „other" consequences are so severe that a mere focus on the political system would fall short of the mark. This applies to the environment, work, education, society and law. The enormous hunger for energy and the associated CO_2 emissions of data centres for AI, as well as their cooling water and land consumption and copyright issues, should be highlighted here as examples.

[5] „Other" consequences no less significant

Against the background of analysing the opportunities and risks of Generative AI concerning the democratic political system, the study concluded by asking whether society can passively observe and allow the socio-technical development that has been initiated or whether it would be more appropriate to shape the process, i.e. attempt to intervene in one way or another. To answer this question, a broad search was carried out for options for action that would promote desirable directions of development and the exclusion or at least minimisation of recognised risks. This study distinguishes between regulatory, organisational and technical approaches. These numerous options, found in scientific and popular literature, policy papers, legal acts, and the media, were structured and analysed in a multi-stage process and finally condensed for the primary addressee, the Austrian parliament. In this way, the following key conclusions were drawn:

[6] Compilation and analysis of the options for action:
– regulative
– organisational
– technical

[7] Conclusions for the Parliament:

Strengthening democratic discourse

- Set up a parliamentary commission of enquiry on „Democracy and AI"
- Develop a code of conduct on AI in politics
- Hold nationwide citizens' forums on fundamental questions of democracy
- Massively promote media and AI literacy
- Foster transparency through labelling and promotion of approaches to increase explainability
- Discuss specific regulatory proposals, especially:
  - Define the responsibility of platforms for content
  - Enact a general ban on political microtargeting
  - Ban politically motivated deepfakes meant to cause harm

Building digital sovereignty on democratic principles

- Position Austria in the EU as a pioneer of a proactive approach to the preservation of democracy
- Set initiatives towards state digital sovereignty
- Building a European discourse platform that is compatible with democracy
- Systematically and consistently fend off attempts at manipulation, especially:
  - Promote fact-checking
  - Intensify coordination against hostile interference
  - Enforce comprehensive labelling

Exploring opportunities, promoting accompanying research

- Develop promising AI applications in a political context
- Accelerate accompanying AI research
- Annual monitoring report on digital political communication in Austria

A mix of options is necessary

This abundance of options shows that regulation alone is not sufficient when dealing with the risks caused by Generative AI. Instead, it takes a combination of different, well-coordinated measures – regulatory, organisational, and technical – to effectively counter the threat to democracy.

VORWORT

Bezugspunkt „Demokratie“

Im Titel dieser Studie steckt zentral das Wort „Demokratie“. Österreich ist eine demokratische Republik (Art. 1 B-VG), die Europäische Union gründet sich auf den Wert der Demokratie (Art. 2 EU-V). Wie Demokratie formal funktionieren soll, wird in der österreichischen Bundesverfassung ebenso wie in den EU-Verträgen in vielen Bestimmungen ausbuchstabiert. Jenseits dieser formalen Festlegungen gibt es die politisch-demokratische, sich stetig wandelnde Praxis. Expert:innen der Rechts-, Staats-, Politikwissenschaft, der Philosophie und letztlich die Staatsbürger:innen diskutieren daher seit jeher darüber, was demokratischen Grundsätzen entspricht oder nicht. Es gibt dementsprechend auch verschiedene Demokratie-Modelle und analytische Kategorien: von repräsentativen Modellen über direktdemokratische und Mischformen, von Konkurrenz- und Konkordanz-, Mehrheits- oder Konsensdemokratie bis zu sozialistischer oder liquider Demokratie. Schließlich werden von manchen bestimmte Erscheinungsformen als defekte Demokratie bezeichnet, etwa die gelenkte oder die „illiberale“ Demokratie, und es gibt den Befund der Postdemokratie. Wenn die vorliegende Studie also auf den Einfluss einer neuen Technologie (der sog. Generativen Künstlichen Intelligenz) auf etwas so Komplexes wie die Demokratie fokussiert, sollte vorab klargestellt werden, worauf wir uns beziehen. Angesichts der angedeuteten Vielfalt demokratischer Formen, der inhärenten Dynamik der staatlichen Entwicklung und der enorm umfangreichen Literatur zum Demokratiebegriff kann hier freilich keine umfassende Exegese geleistet werden.

Die Demokratie österreichischer Prägung

Der Bezugspunkt dieser Studie ist naheliegenderweise die Demokratie österreichischer Prägung im europäischen Kontext. Österreich ist eine liberale Demokratie. Dazu gehören allgemeine, freie und geheime Wahlen, die Garantie der Grundrechte sowie die Gewaltentrennung (Legislative, Exekutive, Judikative). Österreich hat ein pluralistisches politisches System, indem eine Vielfalt von gesellschaftlichen Kräften respektiert und die Macht prinzipiell auf verschiedene Institutionen verteilt wird. Es ist eine repräsentative Demokratie, in der gewählte Mandatar:innen (Abgeordnete zum Parlament, aber auch Funktionsträger:innen auf Länder- und Gemeindeebene) die zentralen politischen Entscheidungen treffen. Politische Parteien haben maßgeblichen Anteil an dieser Willensbildung. Aufgrund der Tatsache, dass in aller Regel die Regierung eine Mehrheit im Parlament hat, spielen die Regierungsparteien in der praktischen Willensbildung die entscheidende Rolle. Wichtige weitere Elemente der österreichischen Demokratie sind die parlamentarische Opposition und deren Kontrollrechte, das Prinzip der freien Meinungsäußerung samt Pressefreiheit, der Minderheitenschutz und die Möglichkeit friedlicher Regierungswechsel. Direktdemokratische Elemente sind in Österreich hingegen wenig ausgeprägt und spielen nur eine marginale Rolle. Österreich ist ein Rechtsstaat, der in der Lage ist, diese demokratische Ordnung zu schützen, etwa durch unabhängige Verfassungsgerichtsbarkeit, Strafgerichtsbarkeit und Korruptionsstaatsanwaltschaft. Österreich ist Teil des verfassungsrechtlich verankerten, europäischen Grundrechtsschutzsystems (EU und Europarat). In den international vergleichenden Demokratie-Rankings (The Economist, V-Dem etc.) gilt Österreich als „full (liberal, representative) democracy“, wenngleich nicht auf den vordersten Plätzen weltweit.

Der freie öffentliche Diskurs als zentrales Merkmal einer funktionierenden Demokratie

Im Zusammenhang mit dieser Studie ist vor allem die demokratische Qualität des öffentlichen Diskurses zentral. Dieser hat in zweierlei Hinsicht eine besondere Bedeutung für die demokratische Willensbildung und damit die Demokratie: Einerseits hängen von Art und Verlauf des Diskurses, den diskutierten Inhalten und dem Vertrauen in die dabei benutzten Medien direkt die Entscheidungen der Wähler:innen an den Urnen ab. Aber auch Umfang und Art der Ausübung demokratischer Rechte wie des Versammlungs- und Demonstrationsrechts, wird davon direkt beeinflusst. Andererseits hat der öffentliche (bzw. veröffentlichte) Diskurs indirekt Auswirkungen auf die Entscheidungsfindung innerhalb des politischen Systems, da die Entscheidungsträger:innen darauf Bezug nehmen und sich neben mannigfaltigem Input von Expert:innen und Stakeholdern maßgeblich daran orientieren, was die Bürger:innen zu wollen scheinen bzw. die Meinungsführer:innen argumentieren. Es dürfte in diesem Sinne unbestritten sein, dass es für das demokratische System Österreichs somit entscheidend ist, dass der öffentliche Diskurs von höchstmöglicher Qualität ist. Damit ist der Idealzustand gemeint, dass die diversen Foren der staatsbürgerlichen Meinungsbildung potenziell allen Staatsbürger:innen alle relevanten Informationen zur Verfügung stellen und vor unlauterer Einflussnahme und Verzerrung geschützt sind. Damit eine im Sinne der Verfassung „freie Wahl" garantiert werden kann, muss die Willensbildung frei und unbeeinflusst sein. Die Staatsbürger:innen sollten das Vertrauen haben, dass gerade im öffentlichen Diskurs Regeln eingehalten werden, die diese Freiheit garantieren. Dazu zählt in erster Linie das Recht auf Meinungsäußerung, aber auch die Einhaltung von deren Grenzen, insbesondere der Schutz vor persönlicher Verunglimpfung und Betrug, die Einhaltung der Gesetze gegen politische Delikte usw. Die Grenze zwischen dem, was zulässig ist und was nicht, ist oft schwer zu ziehen und beschäftigt schon heute Presseräte und Gerichte.

Neue Herausforderungen in der digitalen Welt

Gerade im Zusammenhang mit der teilweisen Verlagerung des öffentlichen Diskurses in die sog. Sozialen Medien, aber auch in die Diskussionsforen der traditionellen Medien, die längst auch im digitalen Raum präsent sind, zeigt sich, dass sich für den Erhalt der freien Willensbildung der Staatsbürger:innen gänzlich neue Herausforderungen stellen. Schon in den letzten Jahren war von den sog. Filterblasen oder Echokammern die Rede, von Microtargeting, Fake News und hybrider Kriegsführung mittels digitaler Propaganda. Nun sind gerade erst weitere Herausforderungen für den öffentlichen Diskurs durch Anwendungen Generativer KI sichtbar geworden, die die frühere Herausforderung noch verstärken und teils ganz neue Qualitäten zeigen.

Was die Technikfolgen-abschätzung leistet

Technikfolgenabschätzung (TA) basiert auf der Grundlage solider interdisziplinärer wissenschaftlicher Forschung und soll einen Beitrag zu gut informierten politischen Entscheidungen leisten, die auf dem Stand des Wissens und der Analyse der Unsicherheiten des technischen Wandels basieren. Die TA bemüht sich um Äquidistanz zu allen Parteien und um einen ausgewogenen Blick auf die Partikularinteressen und Bestrebungen von Interessengruppen (Unparteilichkeit) – diese Studie ist dementsprechend auch von allen Parlamentsfraktionen konsensual beauftragt und richtet sich wiederum an alle im Parlament vertretenen Parteien (und die interessierte Öffentlichkeit). Die hier diskutierten Themen werden aus allen potenziellen Perspektiven betrachtet (Multiperspektivität) – nicht zuletzt durch die Einbeziehung von Meinungen und Wissensbeständen mittels Interviews und in einem Workshop; Werte, Interessen und implizites Wissen werden explizit gemacht (Transparenz) – etwa, wie im Vorstehenden geschehen, in

Hinblick auf die Werte der Demokratie und der Menschenrechte. TA enthält sich prinzipiell normativer Aussagen, ihr Analyserahmen ist jedoch durch die normative Basis der demokratischen Werte geprägt – insb. in Bezug auf den Erhalt des oben beschriebenen, in der Bundesverfassung verankerten liberalen, repräsentativen, rechtsstaatlichen, pluralistischen demokratischen Systems. Im Schlusskapitel werden die möglichen politischen Handlungsoptionen in diesem politisch sensiblen Themenbereich dargestellt. Die vorliegende Studie steht somit in dieser gefestigten Tradition der TA.

1 EINLEITUNG

„Informierte Bürger:innen sind die Grundlage der demokratischen Debatte und Gesellschaft. Die beschleunigte Verbreitung falscher oder irreführender Informationen, oft durch gezielte Desinformationskampagnen in- oder ausländischer Akteure, stiftet Verwirrung und verschärft die Polarisierung, verzerrt öffentliche politische Debatten und verschlechtert das Vertrauen in die Regierung weiter. In einer sich schnell verändernden Informationslandschaft, die durch die Digitalisierung neu geformt wird, sind die Stärkung der Integrität von Informationsräumen und die Bekämpfung von Desinformation daher dringend erforderlich, um das soziale Gefüge offener Gesellschaften und die Demokratie zu stärken."[1]

1.1 DAS THEMA DIESER STUDIE

Demokratien im ständigen Wandel …

Demokratische Systeme befinden sich kontinuierlich im Wandel. Die Zusammensetzung des politischen Personals verändert sich über die Zeit aufgrund neuer Mehrheitsverhältnisse und Generationenwechsel, wodurch die von Einzelnen (mit-)geprägte politische Kultur weiterentwickelt wird. Die inhaltlichen Anforderungen an die demokratische Politik unterliegen zahlreichen dynamischen Faktoren, darunter die sich verändernde Demographie (Alterspyramide, Migration), neue Herausforderungen der natürlichen Umwelt (wie der Klimawandel), wirtschaftliche Entwicklungen, Kriege und sich ändernde geopolitische Machtverhältnisse. Rechtliche Dynamiken spielen ebenfalls eine Rolle, angefangen auf der supranationalen Ebene, die die demokratischen Spielräume auf nationaler Ebene beeinflussen, bis hin zur Judikatur der Verfassungsgerichte.

…nicht zuletzt durch Technik …

Technik ist ein weiterer bedeutender Faktor für gesellschaftlichen und demokratischen Wandel. Obwohl Wahlen vielerorts noch auf Papier stattfinden und der Parlamentsbetrieb größtenteils in mündlicher und schriftlicher Form abläuft oder es weiterhin auch gedruckte Zeitungen gibt, ist der zunehmende Einsatz neuer Technologien nicht zu übersehen. Dies zeigt sich sowohl innerhalb als auch außerhalb der politischen Institutionen. Insbesondere die Digitalisierung hat auch die Einrichtungen der Demokratie längst erreicht: Von der Außendarstellung im Internet und Kommunikationsflüssen im Intranet aller politischer Institutionen über die Bedeutung der Sozialen Medien für politische Meinungsbildung, insbesondere in Vorwahlzeiten, und für den politischen Diskurs bis hin zur Live-Übertragung von Parteiveranstaltungen oder Parlamentsdebatten. Informations- und Kommunikationstechnik spielt somit eine zunehmend wichtige Rolle im demokratischen Prozess. Karaboga et al. (2024, S. 264ff) beschreiben dies als „digitalen Strukturwandel der Öffentlichkeit". Matasick et al. (2024, Kap. 1.2) sprechen von „changes in information spaces affecting democratic engagement". Das Thema

…insbesondere Informations- und Kommunikationstechnik …

[1] Zitat aus dem Vorwort des aktuellen Berichts für die OECD von Matasick et al. (2024); Übersetzung aus dem englischen Original mit Hilfe von deepL.com.

„Digitalisierung und Demokratie" ist bereits vielfach Gegenstand von internationalen Konferenzen (z.B. Bogner et al. 2022; EPTA 2023).[2]

… und jüngst sogenannte Künstliche Intelligenz …

Im Kontext des fortwährenden Wandels stellen die Anwendungen sogenannter Künstlicher Intelligenz (KI) eine besondere Entwicklung dar (vgl. dazu etwa die US-Studie mit weltweitem Vergleich „Artificial Intelligence and Democratic Values 2022", Caunes 2023).[3]

… insbesondere Generative KI (Definition)

Obwohl KI bereits seit Jahrzehnten Gegenstand von Forschung und Entwicklung ist und in verschiedenen Bereichen, wie beispielsweise in Medizin und Forschung, beträchtliche Erfolge erzielt hat, ist erst kürzlich durch den Aufstieg der sogenannten Generativen KI eine breitere Aufmerksamkeit entstanden. Der Begriff „Generative KI" bezieht sich auf *Software, die mithilfe von Maschinellem Lernen und umfangreichen Datensätzen (Texte, Bilder, Audio) unter Einsatz statistischer Methoden und neuronaler Netze neue Inhalte generieren kann*.[4]

Das aktuell bekannteste Beispiel ist ChatGPT, ein Sprachroboter (Chatbot), der auf Basis eines umfangreichen Sprachmodells (Large Language Model, LLM) mit Nutzer:innen über text- und sprachbasierte Nachrichten und Bilder zu praktisch jedem Thema kommunizieren kann. Die Sprachqualität von ChatGPT ist in einigen Sprachen mittlerweile verblüffend hoch, während die inhaltliche Zuverlässigkeit oft fragwürdig scheint. Es ist jedoch davon auszugehen, dass in diesem äußerst dynamischen Technologiefeld die aktuellen Defizite durch die Kombination von Sprachmodellen mit Expertensystemen und themenspezifischen Konfigurationen tendenziell minimiert werden können. Es gibt allerdings auch die Vermutung, dass dieser Verbesserung aufgrund der Abhängigkeit von Trainingsdaten Grenzen gesetzt sind bzw. es sogar wieder zu Verschlechterungen kommen könnte. Neben Sprachrobotern gibt es zahlreiche weitere Anwendungen zur künstlichen Erzeugung neuer Texte, Übersetzungen, Bilder, Videos, Musik, Sprache und Computercode. Insbesondere bei audiovisuellen Inhalten ist es für Menschen zunehmend schwierig bis unmöglich, zwischen menschengemachten und KI-erzeugten Inhalten zu unterscheiden – etwa bei sogenannten Deepfakes oder der subtilen Beeinflussung durch Chatbots.[5]

Risiken für den öffentlichen Diskurs durch Desinformation und Deepfakes

Generative KI-Anwendungen verbreiten sich derzeit rasant aufgrund ihrer Leistbarkeit, einfachen Bedienbarkeit und schnellen Generierung von Inhalten. Diese Verbreitung erstreckt sich sowohl auf die allgemeine Bevölkerung als auch auf spezialisierte Anwendungen, insbesondere im Medienbereich,[6] und beginnt, den öffentlichen Diskursraum zu transformieren. Dies birgt nicht zuletzt für die

---

2 NTA9-TA21, ***oeaw.ac.at/ita/veranstaltungen/vergangene-veranstaltungen/konferenzen/nta9-ta21-konferenz***, EPTA'23, ***parlament.cat/pcat/epta-2023/***.
Anmerkung: Alle in diesem Bericht zitierten Weblinks (URLs) wurden zuletzt am 21.01.2025 überprüft.

3 Es sei an dieser Stelle erwähnt, dass es nicht nur Diskursverzerrung *durch* KI-Anwendungen gibt – das ist der Fokus dieser Studie – sondern auch *über* KI: Künstliche Intelligenz ist nicht nur technisch sehr mächtig, sondern auch als Hype im Diskurs, gibt es doch sowohl überzogene Heilserwartungen als auch Weltuntergangsbefürchtungen – die jedoch nicht Gegenstand dieser Studie sind.

4 Siehe Abschnitt 2.1 für eine genauere Bestimmung von „Generativer KI".

5 Siehe dazu bereits den FTA-Monitoringtext zu „Deepfakes – Perfekt gefälschte Bilder und Videos" (***parlament.gv.at/fachinfos/rlw/Deepfakes-Perfekt-gefaelschte-Bilder-und-Videos***). Im Abschnitt 2.2 wird der Begriff „Deepfake" genauer spezifiziert.

6 Vgl. bspw. die kürzlich preisgekrönte generative KI-Anwendung des ORF namens „AiDitor", ***orf.at/stories/3360846/***.

Demokratie Chancen wie auch Risiken.[7] Insbesondere der bereits seit längerem zu beobachtende Einsatz von nicht-faktenbasierten manipulativen Inhalten (sog. Fake News) in den Sozialen Medien wird durch Generative KI deutlich einfacher, billiger und – aufgrund der für Menschen höheren Überzeugungskraft von Bewegtbildern – vermutlich noch wirksamer (vgl. Matasick et al. 2024, Kap. 1.2). Ein weiterer massiver Anstieg des Einsatzes von KI-Technologie und deren Auswirkungen ist daher zu erwarten, vor allem weil die hohe Reichweite und die Algorithmen der Sozialen Medien das Verbreiten von Deepfakes begünstigen (Karmasin et al. 2024, S. 23). Die bisherigen Techniken und Kapazitäten zu deren Detektion und bestehende Leitlinien werden dieser Dynamik noch nicht gerecht.[8] Dieser Umstand führt bereits zu lebhaften Debatten im Feuilleton sowie unter Expert:innen[9] und war Ausgangspunkt etlicher Studien: Jüngst beschäftig(t)en sich u. a. auch die Schweizer TA-Einrichtung TA-Swiss auf über 400 Seiten mit diesem Thema (Karaboga et al. 2024, insb. Kapitel 6 "Deepfakes in der Politik") sowie die deutsche parlamentarische TA-Einrichtung TAB (Madeira et al. 2024), aber auch die OECD (Matasick et al. 2024); mit speziellen Aspekten, teils nicht spezifisch für Generative KI, auch das Europäische Parlament/STOA (Dumbrava 2021; Villar García et al. 2021) und die US-amerikanische TA-Einrichtung (STAA 2023).

Scoping: Der Fokus der Studie im Überblick

Es ist wohl unerlässlich, dass innovative Ansätze und verbesserte Strategien entwickelt werden, um den Herausforderungen für die Demokratie, die durch die fortschreitende Verbreitung von Generativer KI entstehen, effektiv zu begegnen. Vor diesem Hintergrund stellten uns die Abgeordneten folgende Frage:

> *„Wie können wir in einer Demokratie mit dieser neuartigen Manipulation von Informationen und potentieller Beeinflussung der öffentlichen Meinung umgehen und wie können wir sie in einem ersten Schritt überhaupt erkennen?“*

Das Parlament wünschte sich somit eine Aufarbeitung der Auswirkungen von Generativer KI und hier insbesondere von Deepfakes auf die Gesellschaft im öffentlichen Bereich, einschließlich der Politik, der Rolle des Parlaments und von Wahlen. Die vorliegende Kurzstudie konzentriert sich in diesem Sinne spezifisch auf Auswirkungen von Generativer KI auf den demokratischen Prozess, also insbesondere den öffentlichen (politischen) Diskurs. Darüber hinaus werden auf Wunsch der Abgeordneten auch die problematischen Aspekte für von Deepfakes betroffenen Einzelpersonen adressiert werden. Dabei konzentriert sich die Studie auf die mögliche Diskreditierung oder sogar Erpressung von politischen Amtsträger:innen. Weiters werden auch die Potenziale von Generativer KI für demokratische Institutionen, insbesondere das Parlament, dargestellt und analysiert. Ein besonderer Schwerpunkt wird auf Mechanismen des Erkennens der Manipulationen gelegt.

---

7 Siehe dazu bereits den FTA-Monitoringtext zu „Generative KI und Demokratie“ (*parlament.gv.at/fachinfos/rlw/Generative-KI-und-Demokratie*). Zu den Chancen, siehe Kapitel 3 in diesem Bericht; zu den Risiken, siehe Kapitel 4.

8 Vgl. dazu Abschnitt 2.3.

9 Beispielsweise: Tom Valovic, *commondreams.org/opinion/ai-chatgpt-technology-and-democracy*; Felix M. Simon et al., *oii.ox.ac.uk/news-events/news/fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown-says-oxford-ai-researcher/*; Steven Feldmann, *tandfonline.com/doi/full/10.1080/00396338.2023.2261260*; Sarah Kreps und Doug Kriner, *journalofdemocracy.org/articles/how-ai-threatens-democracy/*; Laura Hood, *theconversation.com/generative-ai-like-chatgpt-could-help-boost-democracy-if-it-overcomes-key-hurdles-212664*.

TA-Studie, mit Fokus auf das demokratische System

Die vorliegende Studie legt den Fokus auf die Auswirkungen auf das politische Feld, auf das demokratische System (Österreichs). In einem eigenen Kapitel wird jedoch ein kurzer Überblick auf die sonstigen Folgen von Generativer KI gegeben, etwa in Hinblick auf Umwelt, Gesundheit und Soziales. Eine umfassende Betrachtung aller Spielarten von KI und deren (indirekter) Beeinflussung demokratischer Prozesse ist ebenfalls nicht Ziel dieser, der Fokus liegt speziell auf Generativer KI.

Gliederung der Studie

Die Studie exploriert das Thema wie folgt: In einem weiteren Abschnitt der Einleitung werden zunächst zentrale Schlüsselbegriffe definiert, die für das Verständnis der Technologie und ihrer Folgen und damit auch für die Entwicklung von Maßnahmen von Bedeutung sind (1.2). Kapitel 2 fasst den Stand der Technik zusammen: Generative KI allgemein (2.1), Deepfakes (2.2) und Techniken zur Detektion von durch Generative KI erzeugten Inhalten (2.3). Kapitel 3 ist den Potentialen Generativer KI im politischen Bereich, insb. im Parlament, gewidmet. Kapitel 4 und 5 erörtern anschließend die Risiken der neuen Techniken und deren Verbreitung: Das sind einerseits Risiken der Diskursverzerrung, zunächst allgemein (4.1), dann spezifisch durch Deepfakes im Besonderen (4.2) sowie durch Microtargeting (4.3), andererseits durch die Machtkonzentration im digitalen Sektor (4.4). Der Blick auf Hybride Bedrohungen (4.5) und Cyberkriminalität (4.6) rundet das Bild ab und wird in 4.7 mit der Brille der Digitalen Souveränität zusammengefasst. Im Kapitel 5 werden darüber hinaus kursorisch sonstige Wirkungsdimensionen von Generativer KI, die nicht unmittelbar mit Demokratie zu tun haben, vorgestellt. Das Abschlusskapitel enthält einerseits eine Zusammenstellung der im Fachdiskurs und in der Öffentlichkeit auf verschiedenen Ebenen gemachten Vorschläge zum Umgang mit den Herausforderungen Generativer KI in drei Perspektiven: regulative (6.1), organisatorisch-institutionelle (6.2) und technische (6.3) Maßnahmen. Im letzten Kapitel werden konkrete Handlungsoptionen für das österreichische Parlament und andere Akteure vorgestellt (7).

1.2 KLÄRUNG DER SCHLÜSSELBEGRIFFE

Zentraler Bezugspunkt öffentlicher (politischer) Diskurs

In dieser Studie geht es in erster Linie um die Wirkungen, die Generative KI in all ihren Formen (siehe Kapitel 2) im demokratischen System Österreichs (siehe Vorwort) auf den *öffentlichen (politischen) Diskurs* haben kann. Ohne an dieser Stelle die breite wissenschaftliche Literatur zum Begriff des (öffentlichen) Diskurs (z. B. bei Habermas und Foucault) aufzubereiten, wird dieser zentrale Bezugspunkt im Rahmen dieser Studie in einem möglichst breiten Sinne verstanden und reicht von der Kommunikation in Print- und audiovisuellen Medien über Soziale Medien bis zu allen Arten von Diskursen unter Anwesenden, Wahlkampfauftritten, Parlamentsdebatten etc.

Im Zusammenhang mit der Verbreitung generierter, falscher oder irreführender Informationen werden zahlreiche Begriffe verwendet, unter anderem Fake News, Deepfakes, Desinformation und Misinformation. Dieses Begriffe überlappen sich teilweise, bedeuten teilweise dasselbe, haben teilweise eine definierte Bedeutung, werden jedoch vielfach von politischen Akteur:innen unterschiedlich und mit unterschiedlichem normativen Gehalt verwendet.

Da für diese Begriffe in der Regel keine scharfen Legaldefinitionen existieren und eine hinreichende Differenzierung der verschiedenen Begriffe fehlt, werden sie manchmal auch synonym zur Beschreibung der allgemeinen Problematik falscher Informationen verwendet, die von Wardle/Derakhshan (2017) auch als „Information Disorder" bezeichnet wird. Demgegenüber steht der Begriff der „Information Integrity" – definiert als Informationsumgebungen, die die Verfügbarkeit von akkuraten, evidenzbasierten und pluralistischen Informationen fördern und die es Individuen ermöglicht, sich unterschiedlichsten Ideen auszusetzen, Entscheidungen informiert zu treffen und ihre bürgerlichen Rechte besser auszuüben (Matasick et al. 2024, Kap. 1). Nachfolgend sollen die genannten Begriffe gemeinsam mit weiteren zentralen Begriffen im Kontext Generativer KI beschrieben werden, um zu einem möglichst einheitlichen und klaren Verständnis zu gelangen. Einen ersten Überblick gibt die folgende Wordcloud in Abbildung 1; in der Folge werden die für diese Studie wichtigsten Begriffe näher beschrieben und soweit möglich definiert.

Information Disorder versus Information Integrity

Abbildung 1: Wordcloud zu Begriffen rund um das Thema Diskursverzerrung
Quelle: Eigene Darstellung; Schriftgröße durch Autor:innen gewählt

Wir stützen uns in der Folge insbesondere auf die Begriffsklärungen in der TA-Swiss-Studie „Deepfakes und manipulierten Realitäten" (Karaboga et al. 2024, S. 81f) und im OECD-Bericht „Facts, not Fakes" (Matasick et al. 2024, ch.1), in der Mitteilung der Europäischen Kommission (2018) sowie auf die Beiträge von Wardle/Derakhshan (2017); Burghardt (2024); Wardle (2019); Dreyer et al. (2021), Villar García et al. (2021); Fallis (2015); Holzer/Sengl (2020); Tandoc et al. (2018), Miró-Llinares/Aguerri (2023), van Huijstee et al. (2021); Kapantai et al. (2021); sowie Rudnicki/Steiger (2020).[10]

[10] Siehe auch: *rtr.at/medien/was_wir_tun/DigitaleDienste/DSA/Desinformation.html*.

1.2.1 BASISBEGRIFFE

Ohne Anspruch auf Vollständigkeit und autoritative Definitionen werden in diesem Abschnitt einige Begriffe erläutert, anhand derer die Diskussion um generierte oder manipulierte digitale Inhalte in diesem politisch aufgeladenen Themenfeld verständlicher wird und auf die auch die folgenden Analysen dieses Berichts aufbauen.

Echtheit oder *Authentizität* spielt bei der Definition von Deepfakes oder textlichen Zitaten eine wichtige Rolle. Wenn etwa einer Person Worte oder Handlungen mittels Deepfake-Technologien unterstellt werden, welche diese Person nie gesagt oder getan hat, handelt es sich dabei um keinen echten Inhalt, sondern um eine Fälschung. Karaboga et al. definieren knapp: „Mit Echtheit bezeichnen wir, ob eine Tatsache mit ihrer Darstellung, z. B. in Form eines Medieninhalts, übereinstimmt." (2024, S. 76) Echtheit und Originalität sind nicht dasselbe. Als *original* definieren Karaboga et al. einen digitalen Inhalt, „welcher erstmalig als solcher erstellt oder gespeichert wurde". Das bedeutet, dass ein Original sowohl echte als auch nicht-echte Inhalte darstellen kann. Mit der Zuschreibung als „Original" wird damit lediglich ausgesagt, dass der erstmalige Inhalt nicht verändert wurde (S. 76).

Echtheit (Authentizität) und Original

Demgegenüber versteht man unter einer *Fälschung* einen Inhalt, der in Täuschungsabsicht einen echten oder originalen Inhalt nachbildet. Wie diese Nachbildung erfolgt, ist unerheblich, sei es, indem Originalmaterial manipuliert wurde, sei es, dass ein Inhalt überhaupt neu erstellt wurde.[11] Dabei ist die Absicht entscheidend, die Rezipient:innen zu verführen, die Fälschung für das Original oder den echten Inhalt zu halten (Karaboga et al. 2024, S. 76). Auch bei einer *Imitation* handelt es sich um eine Nachbildung der Eigenschaften eines Originals, jedoch in der Regel ohne Täuschungsabsicht (S. 77), etwa in der darstellenden Kunst.

Fälschung vs. Imitation

Wenn ein originaler Inhalt verändert wird, also nicht mehr dem Original entspricht, spricht man von *Manipulation*, was in der Regel mit Täuschungsabsicht einhergeht, aber nicht immer, denn irreführende Inhalte, also Text-, Audio-, Bild- und Videoinhalte, die den Betrachter:innen als authentisch/echt erscheinen, es aber nicht sind, können auch ohne Irreführungsabsicht entstehen. Wie die Manipulation durchgeführt wird, ist unerheblich, denn das geht natürlich auch ohne KI, etwa durch das Herausschaben von Personen aus Fotos. Zweifellos finden diese Veränderungen heute jedoch vorwiegend mit digitalen Tools statt, die ihrerseits zunehmend KI-unterstützt arbeiten. Dadurch werden die Rezipient:innen getäuscht bzw. in die Irre geführt. „Eine Irreführung kann also sowohl von einem bewusst zur Irreführung produzierten Inhalt ausgehen als auch von Inhalten, die z. B. zu Unterhaltungszwecken produziert wurden." (S. 77) Insofern ergeben sich aus dem Oberbegriff „Irreführung" die beiden Unterbegriffe Desinformation bzw. Misinformation (siehe nächster Abschnitt 1.2.2).

Manipulation

Manipulierte Medieninhalte

Irreführende Inhalte

[11] Wird der neue Medieninhalt mittels KI-Technologien erzeugt (z. B. durch Eingabe von Textbefehlen in natürlicher Sprache (Prompts) in Software wie ChatGPT oder Dall-E), spricht man von einem *synthetischen Medieninhalt*. Dabei kann es sich um Text-, Audio-, Bild- oder Videoinhalte handeln (vgl. Karaboga et al. 2024, S. 77). Synthetische Medien können auch virtuelle Inhalte in Umgebungen von Virtueller oder Augmentierter Realität (VR, AR) einschließen.

Deutlich schwieriger zu beschreiben, geschweige denn zu definieren sind einige weitere, in der Alltagssprache vielfach, aber je nach Kontext unterschiedlich verwendete Begriffe.

Wahrheit vs. Lüge vs. „Halluzination“

So wird zwischen *Wahrheit*, bzw. einer wahren Aussage, und *Lüge* unterschieden. Am einfachsten wäre es, wenn, was wahr ist, objektiv feststellbar wäre, was aber in der Praxis schwierig ist – außer in banalen Fällen, wenn es etwa darum geht festzustellen, ob es gerade regnet. Philosophisch gesehen gibt es keine absolute Wahrheit, sondern eine Vielzahl subjektiver und intersubjektiver Wahrheiten; erst die geteilten Annahmen einer Gesellschaft sind die Grundlage für die Konstruktion der gemeinsamen Realität. Durch eine Lüge wird absichtlich die Unwahrheit mitgeteilt. Wenn unabsichtlich nicht die Wahrheit gesprochen wird, kann das entweder aus Unwissen geschehen oder weil – im Falle von KI-Textgeneratoren – die Technik so gestaltet ist, dass es nicht ausgeschlossen ist, dass Fakten sozusagen erfunden werden, wenn sie wahrscheinliche Sprachausgaben darstellen und keine inhaltlich passenden Muster in den Trainingsdaten vorliegen (sog. Halluzinationen, siehe dazu Abschnitt 4.1.2).[12]

Erkennen von „Unwahrheit“

In der Praxis wird weniger nach „der Wahrheit“ gesucht als vielmehr versucht, Unwahrheiten zu erkennen. Die Gerichtsbarkeit versucht etwa, aufgrund aufwändiger Beweisverfahren und Befragungen Lügen, also die absichtliche Verdrehung von Tatsachen, von wahren Aussagen zu unterscheiden. Dasselbe gilt etwa für den (Qualitäts-)Journalismus, der ebenfalls durch aufwändige Verfahren versucht, keine Lügen, sondern wahre Aussagen zu verbreiten („Check, Re-Check, Double-Check“). Auch in der klassischen Naturwissenschaft wird im Popperschen Sinne[13] das Augenmerk darauf gelegt zu falsifizieren, also Hypothesen anhand von Experimenten zu widerlegen – daher ist wissenschaftliches Wissen immer nur vorläufig. Ob das jeweils angewandte Verfahren (in der Rechtsprechung, im Journalismus, in der Wissenschaft, durch den/die Einzelne:n) ausreichend war, wird aufgrund unterschiedlicher Sichtweisen, Interessen und Erfahrungshintergründe in der Regel mehr oder weniger strittig sein. Auch das Medium, über das eine Aussage konsumiert wird spielt für deren Überzeugungskraft mitunter eine wichtige Rolle (etwa Papier vs. Digital, Text vs. Fotografie).

Fakten vs. Hypothesen vs. Meinung

Allgemein gesprochen muss eine wahre Aussage logisch widerspruchsfrei sein und enthält überprüfbare *Tatsachen (Fakten)*. Das sind Sachverhalte, die nachweisbar oder zumindest gesellschaftlich anerkannt sind. Dabei ist der Referenzrahmen entscheidend, also in welchem Kontext etwas als Faktum anerkannt wird. Im wissenschaftlichen Sinne wird etwas nur dann als (objektives) Faktum angesehen, wenn dieses empirisch belegbar ist, etwa durch anerkannte Beobachtungsmethoden, Experimente usw. Solange der Beweis nicht erbracht ist, handelt es sich nicht um ein Faktum, sondern um eine *Hypothese* über die Verfasstheit der Wirklichkeit. Die bloße *Meinung*säußerung oder ein reines Werturteil stellt als Mitteilung subjektiver Wertungen den Gegenbegriff zum Tatsachenbegriff dar. Eine Meinung ist also in der Regel eine Bewertung von Fakten oder Artikulation einer Überzeugung. Vielfach wird in der Alltagssprache das Wort Meinung auch für eine nicht überprüfte oder nicht überprüfbare Behauptung verwendet.

---

[12] Über KI-Textgeneratoren und der Anspruch auf Wahrheit, siehe Burghardt (2024, insb. S. 424f).

[13] Wobei bei Popper freilich viele weitere Kriterien heutiger Wissenschaft nicht im Fokus stehen, wie sie insbesondere in den Sozial- und Geisteswissenschaften Praxis sind, die ja i.d.R. nicht experimentell arbeiten.

„Alternative Fakten“

Seit 2017[14] wurde in der Politik (v. a. der US-amerikanischen) vielfach der Begriff der *alternativen Fakten* verwendet, der aus wissenschaftlicher Sicht ein Widerspruch ist, denn es kann zu ein und demselben Sachverhalt nicht gleichzeitig, am selben Ort einen alternativen Sachverhalt geben (etwa unterschiedliche viele Teilnehmer:innen derselben Veranstaltung). Vielmehr könnten je nach Methode der Feststellung eines Sachverhalts unterschiedliche Ergebnisse herauskommen; damit geht es nicht darum, dass anzuerkennen wäre, dass es zwei alternative Fakten, sprich Wirklichkeiten, gäbe, sondern darum, dass keine Einigkeit über die anzuwendende Methode erzielt wurde (wissenschaftlich oder nicht-wissenschaftlich). Der Begriff der alternativen Fakten wird heute zumeist für Positionen gebraucht, die nicht dem breiten Konsens einer wissenschaftlichen Gemeinschaft entsprechen. In aller Regel geht es jedoch, jenseits aller Fakten, um die (politische) Interpretation eines Sachverhalts. Vielfach werden die „Fakten“ des jeweiligen politischen Gegners bzw. der politisch anders ausgerichteten Medien aus rein strategischen Gründen einfach als Fake News bezeichnet (siehe Abschnitt 1.2.2).

Alternativmedien

Ebenfalls verbreitet ist der Begriff der *alternativen Medien* oder Alternativmedien. Damit werden Medien bezeichnet, die sich in irgendeiner Weise von etablierten Medien bzw. einem postulierten Mainstream unterscheiden. Streng genommen ist immer in Fluss, was in einer Gesellschaft Mainstream ist, und die Mediennutzung – und damit eben der Mainstream – ändern sich heute gerade durch die neuen Technologien schneller als zuvor. Der Begriff ist in der Medienwissenschaft nicht trennscharf definiert und wurde zu verschiedenen Zeiten unterschiedlich verwendet. Bis in die frühen 2000er Jahre wurde der Begriff vor allem für politisch links orientierte Medien verwendet, denen es darum ging, sog. Gegenöffentlichkeit herzustellen (sog. Alternativzeitschriften, z. B. Straßenzeitungen, Antiglobalisierungsbewegung etc.). Heute werden damit vielfach, aber nicht nur, Medien aller Art (Zeitschriften, TV-Sender, Internetplattformen), die dem rechtspopulistischen und dem rechtskonservativen politischen Spektrum zugeordnet werden können, so bezeichnet bzw. bezeichnen sich diese Medien selbst so (Paulitsch 2024).[15]

---

[14] Vgl. *de.wikipedia.org/wiki/Alternative_Fakten*.

[15] Für einen Überblick über Alternativmedien, deren unterschiedliche Ausformungen, einschließlich der Bezüge zu Verschwörungstheorien siehe *de.wikipedia.org/wiki/Alternativmedien*, mit weiteren Nachweisen.

1.2.2 SCHLÜSSELBEGRIFF DESINFORMATION UND SEIN UMFELD

Einleitend eine Überblicksgraphik, die die Schlüsselbegriffe Desinformation und Misinformation sowie Beispiele dafür zu systematisieren versucht:

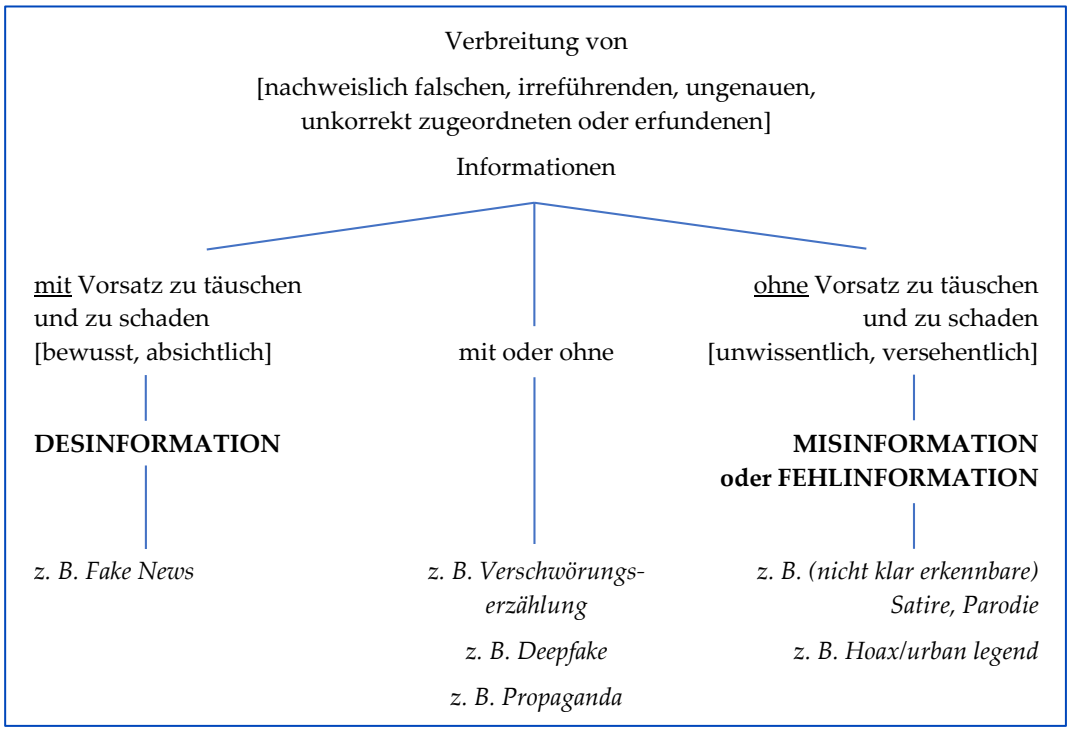


Abbildung 2: Des- versus Mis-Information
Quelle: Eigene Darstellung

Desinformation

Zumeist wird Desinformation (engl. *dis*information), im Rückgriff auf die Begriffsbestimmung der „High Level Expert Group on Fake News and Disinformation" der Europäischen Kommission, definiert als nachweislich falsche oder irreführende Informationen (1), die mit dem Ziel des wirtschaftlichen Gewinns oder der vorsätzlichen Täuschung der Öffentlichkeit (2) konzipiert, vorgelegt und verbreitet werden (3) und öffentlichen Schaden anrichten können (4) (Europäische Kommission 2018, S. 82; Karaboga et al. 2024, S. 77).

Desinformation ist demnach die *intendierte Verbreitung unwahrer oder irreführender Informationen* (so auch Villar García et al. 2021, S. 1; Matasick et al. 2024, Kap. 1).[16] *Einen Grenzbereich bilden hingegen Fälle, in denen Äußernde zwar um die Falschheit der Information wissen und diese dennoch verbreiten, dabei aber entweder keine Täuschungsabsicht haben oder der verbreitete Inhalt kein Täuschungspotenzial aufweist* (Dreyer et al. 2021, S. 6). Äußerungen mit Täuschungsabsicht aber ohne Täuschungspotenzial werden somit nicht als Desinformation verstanden, da sie letztlich auch nicht irreführend sind. Der Aspekt des Täuschungspotenzials einer Äußerung unabhängig von der subjektiven Täuschungsabsicht hat sich aber nicht

[16] Siehe schon die Mitteilung der EU-Kommission aus 2018 „Bekämpfung von Desinformation im Internet: ein europäisches Konzept", *eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236*.

als Bestandteil der Definition von Desinformation durchgesetzt (Dreyer et al. 2021, S. 7). Es werden weiterhin auch unterschiedliche Motive der Äußernden von Falschinformationen unterschieden: So können ideologische bzw. politische Motive dem Verbreiten von Desinformation zugrunde liegen. Manchmal handelt sich es dagegen vor allem um politische Motive bzw. die Absicht des Unruhestiftens („Trolling") oder um wirtschaftliche Motive, etwa bei monetarisierbaren Beiträgen.

Desinformation ist kein neues Phänomen. Mit der digitalen Transformation öffentlicher Kommunikation und ihrer Foren verändern sich aber auch die Erscheinungsformen, die Reichweite bzw. Sichtbarkeit sowie die Wirkungskontexte und -arten von Desinformation. Relevant und aktuell sind vor allem die neuen technischen Möglichkeiten der professionellen Erstellung, der schnellen und automatisierten Verbreitung sowie der beobachtbare hohe Interaktionsgrad von Rezipient:innen. Es zeigt sich, dass nicht nur der Inhalt einer solchen Nachricht selbst, sondern gerade die Kombination aus Inhalt und großen Reichweiten individuell und gesellschaftlich relevante Risikopotenziale aufweist (Humprecht, 2019, 1973 in: Dreyer et al. 2021).

Zentraler Aspekt des Phänomens der problematischen Informationen (Desinformation) ist im Allgemeinen, dass die übermittelte Information ungenau, irreführend, inkorrekt zugeordnet oder einfach falsch bzw. erfunden ist (vgl. Abbildung 2). Je nach Intention der äußernden Person, der Gestaltung des Inhalts und der Wahl der technischen Mittel zur Erhöhung der Sichtbarkeit oder Reichweiten, differenzieren sich die Formen problematischer Information weiter aus.

Malinformation

Matasick et al. (2024, Kap. 1) und Wardle/Derakhshan (2017) grenzen von *Des*information noch *Mal*information ab, die wie erstere mit Schädigungsabsicht verbreitet wird, aber nicht falsch, sondern richtig ist. Beispiele dafür sind, dass private Informationen in der Öffentlichkeit verbreitet werden, oder die bewusst irreführende Kontextualisierung von wahren Informationen, etwa im Bereich Klimawandel.

Misinformation bzw. Fehlinformation

Fehlinformation (oder: Misinformation) sind dagegen falsche oder irreführende Inhalte, die *ohne vorsätzliche Schädigungsabsicht* aus Versehen oder unwissentlich weitergegeben werden, deren Auswirkungen jedoch trotzdem schädlich sein können (Karaboga et al. 2024, S. 77; Matasick et al. 2024, Kap. 1). Beispiele dafür sind Fehler bei der Recherche, unklare Sachlagen oder Zweifelsfälle, ungenaue oder missverständliche Formulierungen aber auch spielerische, witzige oder ironische Äußerungen, die satirisch gemeint sind oder einen „Hoax" („urban legend") in die Welt setzen. Die Äußernden wissen zwar um die Unwahrheit der Information, benutzten diese aber als Ausdruck einer Gesellschaftskritik, ohne explizite Täuschungsabsicht (Dreyer et al. S. 6). Das schließt bei den Rezipient:innen aber ein Täuschungspotential nicht aus, so dass die Grenze zwischen freier Meinungsäußerung und Irreführung oft nicht scharf gezogen werden kann. Auch Verschwörungserzählungen sind nicht (unbedingt) an eine Täuschungsabsicht gebunden. Die Äußernden sind oftmals aus subjektiver Sicht von der Wahrheit der verbreiteten Informationen überzeugt und wollen andere davon überzeugen (Fallis 2015, S. 401,411).

Hoax, urban legend, Satire, Parodie

Verschwörungserzählungen

Quelle: *br.de/nachrichten/netzwelt/deepfake-von-merz-was-bedeutet-er-fuer-den-bundestagswahlkampf,UUCtkLv*

Im aktuellen Bundestagswahlkampf in Deutschland teilte der SPD-Abgeordneter Bengt Bergt ein Deepfake-Video von CDU-Kanzlerkandidat Friedrich Merz. Darin sagt Merz an die Zuschauenden gerichtet: „Die CDU/CSU verachtet Sie". Auch das Banner „Achtung: Künstliche Inkompetenz" deutet darauf hin, dass es sich hierbei um Satire handelt.

Box 1: Satirisches Deepfake im deutschen Bundestagswahlkampf

Fake News

Die Gestaltung einer Information hängt auch mit der Zuschreibung von Wahrhaftigkeit zusammen. Beispielsweise erscheinen journalistische Äußerungen für viele Rezipient:innen vertrauensvoller bzw. glaubhafter, weil sie davon ausgehen, dass der Information aus diesen Quellen ein besonderer Wahrheitsanspruch innewohnt (Holzer/Sengl 2020, S. 161ff). Dieser (vermeintliche) Wahrheitsanspruch führt zu einem besonderen Vertrauensverhältnis zwischen der äußernden und der rezipierenden Person – und kann die Wirkung von Täuschungen verstärken. Die vorsätzliche Herstellung und Verbreitung *irreführender oder falscher Informationen, die wie Aussagen mit besonderem Wahrheitsanspruch gestaltet sind*, werden oft unter dem Stichwort „Fake News" diskutiert (Tandoc et al. 2018). Der österreichische Aktionsplan Deepfakes 2022 definiert so:

> *„Fake News sind in den Medien und im Internet, besonders in sozialen Netzwerken, in manipulativer Absicht verbreitete Falschmeldungen. In Abgrenzung zur Desinformation geht es hier um jede Nachricht, auch solche, die ohne Absicht, ein gewisses Ziel zu erreichen, verbreitet werden (siehe unten Desinformation). Deepfakes fallen unter den Begriff der Fake News, wenn sie mit der Absicht hergestellt werden, zu manipulieren. Schon der Begriff Deepfake lässt darauf schließen, dass diese Form der Manipulation unter den Begriff der Fake News fällt."*
> (BMI/BKA/BMEIA/BMJ/BMLV 2022)

Nach allgemeiner Ansicht stellt dieser Begriff eine *Teilmenge von Desinformation* dar. Der Begriff wird vor allem im Kontext der Diskreditierung oder Delegitimierung klassischer Nachrichtenmedien genutzt, die angeblich nicht wahrheitsgemäß oder tendenziös Bericht erstatten. Der Begriff „Fake News" wird in der wissenschaftlichen Debatte inzwischen jedoch vermieden, da er inhaltlich

unscharf und zugleich politisch aufgeladen ist (Wardle/Derakhshan 2017, S. 5; Karaboga et al. 2024, S. 77). So auch Miró-Llinares/Aguerri (2023):

> *„Via a systematic review of the literature, we observe, firstly, that the concept of fake news used in empirical research is limited and should be refocused because it has not been constructed according to scientific criteria and can fail to include relevant elements and actors, such as governments and traditional media."*

Propaganda

Propaganda ist ein Mittel der strategischen Kommunikation, im traditionellen Verständnis insbesondere in totalitären oder autoritären Regimen und richtet sich meistens an die inländische Bevölkerung (Schünemann 2022, S. 34, siehe auch Abschnitt 1.2). Propaganda stellt zielgerichtete Versuche dar, politische Meinungen oder öffentliche Sichtweisen zu formen, etwa durch die Zuspitzung politischer Botschaften, und das Verhalten in eine vom Propagandisten (vor allem staatliche oder politische Akteure) erwünschte Richtung zu steuern. Wenn sich die Propaganda dadurch auszeichnet, dass die Akteure eine bestimmte Agenda auch durch die gezielte Erstellung und Verbreitung von Desinformation, etwa der Manipulation von Erkenntnissen, verfolgen, um die Rezipierenden zu täuschen, so stellt Propaganda eine *Unterform von Desinformation* dar. Charakteristische Aspekte dieser speziellen Form von Desinformation sind hier somit die Herkunft dieser Informationen aus den staatlichen oder politischen Machtbereichen sowie der systematische Einsatz mit dem Zweck, größere Bevölkerungsgruppen nachhaltig in ihren Sichtweisen und Einstellungen zu manipulieren. Insbesondere die Soziale Medien eignen sich als politisches Massenkommunikationsmedium gut für Propaganda (vgl. Strauß 2020, S. 95ff).

Deepfakes

Deepfakes sind mithilfe von KI-Techniken[17] manipulierte oder synthetisierte Bild-, Audio- oder Video-Inhalte, die authentisch erscheinen und in denen eine oder mehrere Personen etwas zu sagen oder zu tun scheinen, was sie nie gesagt oder getan haben (van Huijstee et al. 2021, S. 2; Karaboga et al. 2024, S. 75). Die Definition von Deepfake gemäß Artikel 3 (Begriffsbestimmungen) der europäischen AI-Act lautet folgendermaßen:[18]

> *„ein durch KI erzeugte[r]n oder manipulierte[r] Bild-, Ton- oder Videoinhalt, der wirklichen Personen, Gegenständen, Orten, Einrichtungen oder Ereignissen ähnelt und einer Person fälschlicherweise als echt oder wahrheitsgemäß erscheinen würde."*

Die Bezeichnung Deepfakes ist auf das Pseudonym eines Nutzers der Plattform Reddit zurückzuführen, der 2017 unter dieser Bezeichnung einige der ersten Deepfakes veröffentlichte. Es handelt sich um ein Kofferwort aus ‚Deep Learning' als angewendeter KI-Technologie und ‚Fake'. Nach der Definition des AI-Act fallen darunter sämtliche durch KI erzeugte oder manipulierte audiovisuelle Inhalte, demnach auch Text-to-Image- bzw. Text-to-Video-Generatoren. Es handelt sich dabei nicht um veränderte, sondern um plausible neuartige synthetische Inhalte. Somit sind auch Large Language Models (LLMs) in die Kategorie der Deepfake-Technologien einzuordnen, sofern damit audiovisuelle Inhalte erstellt wer-

---

[17] Wenn die täuschenden Inhalte hingegen mittels Methoden erstellt wurden, die nicht auf KI-Technologien basieren (klassisches „Photoshoppen", das Neu-Zusammenschneiden bestehenden Videomaterials oder das verlangsamte Abspielen eines Videos etc.), spricht man von *Cheapfake* oder *Shallow Fake* (Karaboga et al. 2024, S. 80).

[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, laying down harmonised rules on artificial intelligence and amending Regulations […] (Artificial Intelligence Act, AI-Act), ABl. L vom 12.7.2024, ***eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689***.

den können. Wir sprechen hingegen bei generierten Texten (wie Karaboga et al. 2024, S. 75) von synthetisiertem Text oder KI-generiertem Text, der nicht als Deepfake bezeichnet wird.

Medienmanipulationen

Deepfakes, bei denen realistisch erscheinende Aufnahmen so nicht stattgefundener Ereignisse erzeugt oder manipuliert werden können, stellen eine neuere Form von Medienmanipulation dar. Unter Medienmanipulationen fallen falsifizierbare Aussagen durch die Abweichung von einem objektiv beobachtbaren Zustand. Auch vollständig erfundene Äußerungen und Zuschreibungen fallen in diese Kategorie (Dreyer et al. 2021).

Je nach Nutzung und Kontext können generierte Inhalte einen nützlichen oder schädigenden, einen erwünschten oder unerwünschten Effekt auf bestimmte Menschen, Gruppen oder Organisationen entfalten (siehe Kapitel 3 und 4). Dieser Effekt kann intendiert oder auch nicht intendiert sein. Eine häufige Anwendungsform ist ihr Einsatz zum Zwecke der Verbreitung von Desinformation und Misinformation (siehe oben).

2 STAND DER TECHNIK

In diesem Kapitel wird ein Überblick über den Stand der Technik und deren Funktionsweise gegeben, wobei in 2.1 Generative KI allgemein, in 2.2 Deepfakes im Besonderen sowie in 2.3 technische Hilfsmittel zur Erkennung von KI-generierten Inhalten dargestellt werden.

2.1 GENERATIVE KI

Definition Künstliche Intelligenz

Der Begriff Künstliche Intelligenz (KI) hat zwar bereits eine lange Geschichte, ist aber in der öffentlichen Diskussion nicht klar definiert und selbst in den Wissenschaften gibt es konkurrierende Definitionen (Deutscher Ethikrat 2023). Wir orientieren uns in diesem Bericht an den Begriffsdefinitionen im politischen und regulativen Kontext, u. a. an der Definition der Hochrangigen Expertengruppe für Künstliche Intelligenz (Europäische Kommission 2018) sowie an der Begriffsbestimmung in Artikel 3 der aktuellen Fassung des KI-Gesetzes der EU (AI-Act). Demnach ist ein „KI-System" ein

> *„maschinengestütztes System, das so konzipiert ist, dass es mit unterschiedlichem Grad an Autonomie betrieben werden kann, das nach der Einführung Anpassungsfähigkeit zeigen kann und das für explizite oder implizite Ziele aus den Eingaben, die es erhält, ableitet, wie es Ergebnisse wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugen kann, die physische oder virtuelle Umgebungen beeinflussen können".*[19]

Regelbasiert vs. selbstlernend

Unterschieden werden regelbasierte und lernende KI-Systeme, die aus Daten lernen, wie bestimmte Ziele erreicht werden können. Zu den lernenden Systemen zählt das Maschinelle Lernen (machine learning, ML), wobei man zwischen überwachtem, unüberwachtem und bestärkendem Lernen unterscheidet. Beim überwachten Maschinellen Lernen werden dem System keine Verhaltensregeln vorgegeben, sondern das System lernt anhand sogenannter Trainingsdaten, indem es seine Parameter so anpasst, dass die Abweichung zwischen erwartetem Output und vom System berechneten Output minimiert wird. Arbeiten die Systeme mit neuronalen Netzen, so spricht man von Deep Learning. Während der Trainingsphase werden hierbei die Gewichtungen der Verbindungen des Netzes so angepasst, dass sie der Aussage der verfügbaren Beispiele möglichst nahekommen. Auf die Trainingsphase folgt dann eine Testphase, in der das Verhalten des neuronalen Netzes anhand zuvor nicht gesehener Beispiele überprüft wird, um festzustellen, ob die Aufgabe gut erlernt wurde. Lernverfahren eignen sich für schwer definierbare Aufgaben, die durch symbolische Verhaltensregeln nicht umfassend abgebildet werden können, insbesondere für Wahrnehmungsaufgaben wie Bild- und Spracherkennung (Europäische Kommission 2018). Die Qualität der Ergebnisse ist jedoch stark von Qualität und Umfang der Trainingsdaten abhängig.

Maschinelles Lernen

Neuronale Netze

Deep Learning

[19] *artificialintelligenceact.eu/de/article/3/*.

Output für Menschen nicht direkt verständlich

Die so erzeugten Zusammenhänge zwischen Input und Output sind zumeist hochkomplex, unübersichtlich und für Menschen nicht direkt verständlich oder hinreichend rekonstruierbar. Auf Maschinellem Lernen basierende künstliche Intelligenz unterscheidet sich von herkömmlichen Softwarelösungen folglich dadurch, dass sie Regelmäßigkeiten bzw. darauf aufbauende Regeln selbst erkennt, anstatt vorgegebene Regeln zu folgen. Solche Systeme haben die Fähigkeit, aus den Eingaben letztlich Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen ableiten zu können, die über die grundlegende Datenverarbeitung hinaus geht (siehe auch Erwägungsgrund 12 des AI-Acts).

Definition Generative KI

Von Generative KI spricht man bei einem Deep-Learning-Verfahren, das darauf spezialisiert ist, Texte, Audio- und Video-Dateien oder andere strukturierte Inhalte zu erzeugen (zu generieren). Generative KI-Modelle basieren auf unüberwachtem oder teil-überwachtem Lernen, um große Mengen an Daten von Texten, Bilder oder Videos zu verarbeiten und daraus neue Formen von Inhalten zu erzeugen, die den Originaldaten ähnlich sind, sie aber zumeist nicht direkt wiedergeben. Aus technischer Sicht fußen diese vornehmlich auf zwei unterschiedlichen Architekturen: Generative Adversarial Networks (GANs) und Transformer. Transformer sind insbesondere auf die Verarbeitung und Interpretation sequenzieller Daten spezialisiert und eignen sich daher optimal für Aufgaben der Verarbeitung natürlicher Sprache. Weil Generative KIs oft recht breit eingesetzt werden können bzw. als Grundlage für spezialisierte Anwendungen dienen, werden sie teilweise auch unter dem (umstrittenen[20]) Begriff „Grundlagenmodelle“ (foundation models) diskutiert.

Zwei unterschiedliche Architekturen: GAN und Transformer

Grundlagenmodelle

Transformer-Modelle nutzen bei der Transformation von Input zu Output Datensequenzen statt individueller Datenpunkte. Das bedeutet, Worte werden nicht isoliert verarbeitet, sondern im Kontext der anderen Worte eines Satzes oder Dokuments. Mit einem sog. Attention-Mechanismus kann das Transformer-Modell beispielsweise verschiedenen Wörtern eine unterschiedliche hohe Aufmerksamkeit (engl. attention) zuweisen und so die Aussage des Satzes besser interpretieren. Deshalb bilden Transformer-Modelle die Grundlage großer Sprachmodelle (Large-Language Models – LLMs).[21]

ChatGPT als Beispiel für ein Transformer-Modell

Die Architektur des bekannten ChatGPT beruht beispielsweise auf einem Transformer-Modell. Es zeichnet sich dadurch aus, dass in Texten auch Verbindungen zwischen weit entfernt stehenden sog. Token, also Textbestandteilen, erkannt und eine große Anzahl von Trainingsdaten und Anfragen von großem Umfang verarbeitet werden können. Der erste Schritt der Generierung von Texten mittels GPT besteht in einem weitgehend automatisierten Training mit einer großen Anzahl an Texten. In diesem Schritt werden die sogenannten Parameter festgelegt. Im Laufe der Zeit wurde die Parameteranzahl großer Sprachmodelle gesteigert, um die Qualität zu erhöhen. GPT-4 hat, so wird kolportiert, insgesamt 1,7 Billionen Parameter.[22] Damit einher geht auch die Fähigkeit, zunehmend größere Anfragen zu verarbeiten (um z. B. spezifische Kontextinformation mitzugeben), bei GPT-4 Turbo ist die Kontextgröße z. B. 128.000 Tokens, was etwa 300 Seiten Text entspricht.[23] In einem zweiten Schritt erfolgt ein Feintuning des Systems speziell auf die Reaktion sprachlicher Eingaben (sog. Prompts) für eine spe-

---

20 *hai.stanford.edu/news/reflections-foundation-models*.
21 *alexanderthamm.com/de/blog/generative-ai-eine-uebersicht/*.
22 *the-decoder.com/gpt-4-has-a-trillion-parameters/*.
23 *openai.com/index/new-models-and-developer-products-announced-at-devday/*.

zifische Simulation von Konversationsfähigkeit, Anpassung der Antworten durch menschliches Feedback, um illegale oder unerwünschte Antworten zu unterdrücken (Albrecht 2024). Neuere Versionen sind in der Lage, auch mit gesprochener Sprache und multimodal zu arbeiten, d. h. Inhalte in Text- oder Bildform zu verarbeiten bzw. auszugeben.

Generative Transformer-basierte KI-Modelle können in eine Vielzahl nachgelagerter Systeme oder Anwendungen integriert werden. Durch dieses breite Spektrum an unterschiedlichen Aufgaben werden sie auch als Modelle mit allgemeinen Verwendungszweck gemäß Artikel 3 Nr. 63 des AI-Acts bezeichnet.

GANs als Basis für Deepfakes

Generative Adversarial Networks (GANs) werden insbesondere zur Bildgenerierung eingesetzt und stellen eine wichtige Technologie zur Erstellung von bildbasierten/visuellen Deepfakes dar (siehe nächster Abschnitt 2.2).

Bemerkenswert, aber schwer nachzuvollziehen: Generative KI arbeitet mit Wahrscheinlichkeiten

Wichtig ist zu verstehen, dass diese Systeme nicht deterministisch, sondern stochastisch arbeiten: Es handelt sich um Wahrscheinlichkeiten, die sich aus dem oben beschriebenen Training und mit einer beabsichtigten Randomisierung ableiten, die bestimmen, welcher Output generiert wird. Das heißt auch, dass frei erfundene Textteile ausgegeben werden können (solange die Wortfolgen wahrscheinlich sind), sogar wenn man als Ausgangspunkt des Prompts bestehende Texte angibt, die bspw. zusammengefasst werden sollen. Diese Eigenschaft Generativer KI-Systeme macht sie bis jetzt ungeeignet für Aufgaben, bei denen es vor allem auf valide Aussagen mit einem hohen Maß an Richtigkeit/Gewissheit ankommt, wie bspw. beim wissenschaftlichen Arbeiten, kann aber etwa einen Geschwindigkeitsvorteil und/oder eine Arbeitserleichterung bringen, wenn es um formale Aufgaben im Bereich des Editierens oder Strukturierens geht. Die nichtdeterministische Arbeitsweise der Systeme ist auch in Bezug auf Konzepte zur Erklärbarkeit und Reproduzierbarkeit problematisch – zwei Anforderungen, die jedenfalls an KI-Systeme in der öffentlichen Verwaltung zu stellen wären.

Zukunft?

Generative KI ist noch eine relativ junge Technologie, die in dafür geeigneten Anwendungsbereichen bereits bemerkenswerte Ergebnisse erzielen kann und sich wachsender Verbreitung erfreut. Die Technikentwicklung ist hochgradig dynamisch. In welche Richtung und in welchen Anwendungen wird sich Generative KI voraussichtlich in den nächsten Jahren entwickeln? Der Trend zeigt, dass der Output überzeugender (nicht unbedingt richtiger) und mit immer geringerem Wissen (auf Seiten der Anwender:innen) erstellt werden kann. Gleichzeitig zeigt sich, dass mit einem Bruchteil der Ressourcen kleinere LLMs erstellt werden können, die annähernd ähnliche Qualität liefern. Dadurch ist absehbar, dass die Erzeugung von Inhalten (auch Deepfakes) mittels Generativer KI zunehmen und die Erstellung von Modellen noch weiter verbreitet sein wird. Dadurch kann es einerseits zu einer beschleunigten Weiterentwicklung kommen, die der Motor für neue, innovative Anwendungen sein kann. Andererseits wird es schwieriger werden, die mit Generativer KI verbundenen Risiken (insbesondere mit technischen Mitteln) zu beherrschen (siehe auch Karaboga et al. 2024, S. 122ff).

2.2 DEEPFAKES

In diesem Abschnitt werden die spezifischen KI-Technologien zur Generierung von Bild-, Audio- oder Videoinhalten im Detail beschrieben. Es soll verdeutlicht werden, dass unter dem Dach der Generativen KI neben den im vorangegangenen Kapitel beschriebenen Transformer-basierten Sprachmodellen ein Spektrum an unterschiedlichen Technologien zusammengefasst werden. Um die gesellschaftlichen Folgen des Einsatzes von Generativer KI abschätzen und bewerten zu können, benötigt man einen Einblick über deren Funktionsweise und Grundlagen.

Zur Erstellung von Deepfakes werden unterschiedliche Deep-Learning-Technologien eingesetzt, die ständig weiterentwickelt werden. Man unterscheidet einerseits zwischen Methoden zur Manipulation oder zur Synthese bzw. Generierung von Inhalten, weiterhin zwischen Methoden für unterschiedliche Arten medialer Inhalte (Texte, Audios, Bilder, Videos) und schließlich auch zwischen spezifischen technologischen Verfahren für unterschiedliche Zwecke.

Im Allgemeinen versteht man unter Manipulation audiovisuellen Materials eine mit Täuschungsabsicht verbundene intentionale Veränderung von authentischen Informationen durch Auswahl, Zusätze oder Auslassungen (Forster 2003, vgl. Abschnitt 1.2). Im Kontext von Deepfakes bedient man sich dabei unterschiedlicher Manipulationsweisen wie beispielsweise der Lippensynchronisation, dem Gesichtstausch oder dem Stimmentausch (Kietzmann et al. 2020).

Techniken bildbasierter Deepfakes:

Im Bereich bildbasierter Deepfakes stehen gegenwärtig folgende Techniken zur Verfügung – siehe dazu auch ausführlich das Kapitel „Ist- und Trendanalyse" von Karaboga et al. (2024, S. 86ff.):

Face swapping

- Beim Gesichtstausch (Face swapping) wird ein Gesicht in einer Zieldarstellung durch ein anderes Gesicht ersetzt. Gesichtsausdruck und Hintergrund der Zieldarstellung werden beibehalten. Dies kann auch in Echtzeit erfolgen (Kietzmann et al. 2020; Li et al. 2020). Oft wird diese Technologie eingesetzt, um berühmte Schauspieler:innen in Filmclips einzubauen, in denen sie nie aufgetreten sind, aber auch um (Gesichter von) Personen, zumeist Frauen, in Deepfakes mit pornografischem Inhalt einzufügen (Farid/Schindler 2020).

Facial reenactment

- Bei der Manipulation des Gesichtsausdrucks (Facial reenactment) wird der Gesichtsausdruck einer Zielperson durch den einer anderen Person ersetzt. Es handelt sich um eine identitätserhaltende Technik, d. h. das Gesicht der Zielperson bleibt im Gegensatz zum Face swapping erhalten (Farid/Schindler 2020).

Lip Syncing

- Eine Spielart der Manipulation des Gesichtsausdrucks ist Lippensynchronisierung (Lip Syncing), bei der speziell die Lippenbewegungen manipuliert werden, d. h. die Lippenbewegung einer Person werden auf die Zielperson übertragen. Damit kann der Mundbereich mit einer willkürlichen Audioaufnahme in Einklang gebracht werden (Diakopoulos/Johnson 2019). Der Schauspieler und Regisseur Jordan Peele hat ein besonders überzeugendes Beispiel solch eines Deepfakes produziert, indem er ein Video des ehemaligen US-Präsidenten Barack Obama so verändert hat, dass dieser Präsident Trump beschimpft.[24]

[24] *theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed*.

Face morphing

- Die Verschmelzung von mehreren Gesichtern wird Gesichtsmorphing (Face morphing) genannt. So kann beispielsweise auf kriminelle Art ein Bild erstellt werden, das aus zwei Gesichtern verschmilzt, um authentische Ausweisdokumente zu erstellen, die von mehreren Personen zugleich verwendet werden können (Damer et al. 2018, S. 1).

Full body puppetry

- Beim Ganzkörperpuppenspiel (Full body puppetry) werden die Bewegungen eines Körperteils (Kopf- und Augenbewegungen, Mimik) oder des gesamten Körpers einer Zielperson durch einen Darsteller, der vor der Kamera sitzt, verändert (Chan et al. 2019).

Gesichtsgenerierung

- Mithilfe von Gesichtsgenerierung werden gänzlich neue Gesichter erschaffen, die in der Realität nicht existieren (Synthetische Inhalte).[25]

Voice cloning

Im Bereich von Audio-Inhalten ist insbesondere der Stimmentausch (also das Ersetzen einer Stimme durch eine andere) und die Imitierung und Veränderung von Stimmen (Voice cloning) zu nennen. Die Generierung synthetischer Stimmen (Speech synthesis oder auch Text-to-speech genannt) kommt dann zum Einsatz, wenn keine vorhandene Stimme imitiert werden soll, sondern geschriebene Wörter durch eine Maschine ausgesprochen werden sollen (van Huijstee et al. 2021, S. 12). Mit den aktuellen Text-Generatoren, die auf großen Sprachmodellen beruhen (siehe Abschnitt 2.1), ist neben der Erzeugung von Text auch die Imitation der Schreibweise oder Sprache eines Menschen möglich. Die verschiedenen Generierungs- und Manipulationsformen können zunächst für Bilder oder Sprachinhalte angewendet werden und dann in einem Video kombiniert eingesetzt werden, um einen möglichst authentischen Eindruck zu erzielen. Die Erstellung von Videos ist mittlerweile auch direkt mit Sprachmodellen möglich, bei denen Videos mit Texteingaben generiert werden (Video-Generator oder Text-to-video-Generator). OpenAI stellte beispielsweise das Programm „Voice Engine" zum Klonen von Stimmen vor. Das Modell könne die Stimme eines Menschen auf Basis eines 15-sekündigen Audiooriginals duplizieren.[26]

In der Praxis unterscheidet man folgende, insb. im Bereich Deepfakes eingesetzte KI-Technologien (nach van Huijstee et al. 2021, S. 7):

Autoencoder

- Autoencoder sind selbstlernende KI-Systeme, die rein auf der Ähnlichkeit von Daten beruhen und bei denen die Trainingsdaten keine vordefinierten Kategorien benötigen. Der Encoder wandelt in einem ersten Schritt die Eingabedaten in eine komprimierte Darstellung der wesentlichen Merkmale um, die auch „latente Darstellung" oder „latenter Raum" genannt wird. Der Decoder rekonstruiert dann mithilfe dieser latenten Darstellung eine Ausgabe, die der Eingabe am ehesten entspricht. Durch diesen Prozess kann ein Autoencoder die wichtigen Merkmale der Daten selbst lernen. Mit einem Autoencoder werden beispielsweise Informationen über Gesichtsmerkmale aus Bildern extrahiert, um Bilder mit einem anderen Ausdruck zu erstellen.

Variations-Autoencoder

- Ein Variations-Autoencoder erzeugt eine Wahrscheinlichkeitsverteilung für die verschiedenen Merkmale der Trainingsbilder. Beim Training erstellt der Encoder latente Verteilungen für die verschiedenen Merkmale der Eingabebilder. Da das Modell die Merkmale oder Bilder als Gaußsche Verteilungen und nicht als diskrete Werte lernt, kann es zur Generierung neuer Bilder verwendet werden und ist damit als generatives Modell einzuordnen.[27]

---

25 Z. B. *this-person-does-not-exist.com/de*.

26 *orf.at/stories/3353071/*.

27 *unite.ai/de/what-is-an-autoencoder/*.

– Generative gegnerische Netzwerke (Generative Adversarial Networks, GANs) sind unüberwachte Deep-Learning-Modelle, die sich insbesondere zur Erzeugung neuer Bilder eignen. Dabei werden abstrakte Stilmerkmale von konkreten Merkmalen des Trainingsmaterials im Lernprozess unterschieden und in einer latenten Darstellung ausgegeben. Die Umwandlung in Bilder erfolgt dann mit Hilfe zweier neuronaler Netze. Das erste Netz bezeichnet man „Generator", weil es synthetische Inhalte erzeugt, die den Quelldaten ähnlich sind. Das zweite Netz ist ein sogenannter „Diskriminator", der zwischen echten und generierten Inhalten unterscheiden kann. Das Ergebnis wird dem Generator zugeführt, der durch dieses Feedback eine iterative Optimierung der Qualität vornehmen kann. Gesichtstausch oder Gesichtsmorphing beruhen auf dieser GAN-Technologie (Goodfellow et al. 2014). GANs können auch zur Erstellung von Tonaufnahmen und Text eingesetzt werden (van Huijstee et al. 2021, S. 79).

Generative gegnerische Netzwerke

– Facial reenactment nutzt neben KI-Methoden der Bilderkennung auch Methoden der Computergrafik. Weitere spezielle Verfahren zur Generierung von Bildern sind sogenannte Diffusionsmodelle, die beispielsweise bei Stable Diffusion der Firma Stability AI eingesetzt werden. Diffusionsmodelle sind generative Modelle, die mit Bildern und deren Beschreibungen trainiert werden (bspw: „Eine Katze sitzt auf einem Baum."). Beim Training werden die Bilder zuerst durch Zugabe von sog. „Rauschen" schrittweise bis zur Unkenntlichkeit verfremdet. Anschließend lernt das Modell, wie es dieses Rauschen unterdrücken – also „wegrechnen" – kann, um die ursprünglichen Daten wiederherzustellen. Nach abgeschlossenem Training können dann neue Bilder aus verrauschten Bildern generiert werden.[28]

Computergraphik, Diffusionsmodelle

– In der Praxis erzeugt man mit Stable Diffusion neue Bilder, indem man das gewünschte Motiv in einem Textfeld beschreibt. Das Modell greift dann auf einen dahinterliegenden Datensatz an Bildern und Bildbeschreibungen zu (z. B. Datenbank LAION-5B). Dieser beruht auf einer umfangreicheren Datensammlung, die regelmäßig mit automatisierten, suchmaschinenähnlichen Programmen aus dem Internet gesammelt und archiviert werden. Es werden jedoch nur diejenigen Bilder mit Beschriftung ausgefiltert, die sich maschinell auslesen lassen und eine von den Entwickler:innen definierte Passfähigkeit zwischen Text und Bildinhalt haben. Die Datenbank besteht somit aus selektiertem Text und Bildpaaren, die demnach kein neutrales Abbild der Welt darstellen. Beispielsweise liegt ein Großteil der Bildbeschreibungstexte nur auf Englisch vor (Beuth et al. 2024).

Stable diffusion

– Neben diesen Diffusionsmodellen werden auch Transformer-basierte Text- und Bildgeneratoren angewendet, bei denen unter Eingabe von Textbefehlen neue Inhalte entstehen (z. B. Midjourney, DALL-E). Es handelt sich dabei aber nicht um veränderte, sondern um plausible, neuartige, also synthetische Inhalte, die unter dem Begriff „synthetisierter Text" oder „KI-generierter Inhalt" in der Literatur eingeordnet werden (vgl. Abschnitt 1.2 und Karaboga et al. 2024, S. 77).

Transformer

[28] *unite.ai/de/Diffusionsmodelle-in-der-KI-–-alles,-was-Sie-wissen-müssen/*.

Vor- und Nachteile der unterschiedlichen Techniken

Autoencoder, GANs und Diffusionsmodelle besitzen spezifische Vor- und Nachteile hinsichtlich der Ergebnisqualität und der Effizienz. Diffusionsmodelle können schnell trainiert werden, wohingegen klassische GAN-Modelle größere Datenmengen für das Training benötigen. Wenn das Training eines GAN-Modells abgeschlossen ist, können keine Änderungen mehr am Modell vorgenommen werden. Neue Modelle erfordern daher immer wieder die Neuberechnung, was sowohl Zeit- als auch Geldressourcen verbraucht (Gupta 2020). Die Kosten für das Training hängen hierbei signifikant von der Anzahl der Parameter ab. Das Training eines Sprachmodells (Natural Language Processing) mit 110 Millionen Parametern kostet zwischen 2.500 Dollar und 50.000 US-Dollar pro Durchlauf. Bei 1,5 Milliarden Parametern sind es bis zu 1,600.000 US-Dollar, was bei der Notwendigkeit von mehreren Durchläufen zu Kosten in achtstelliger Höhe führen kann (Sharir et al. 2020, S. 2; Karaboga et al. 2024, S. 88). Zudem haben GAN-Modelle Probleme, die realistischen Bewegungen von Haaren (z. B. bei Kopfbewegungen) darzustellen. Allerdings wird auch an der Überwindung dieser Herausforderungen geforscht (Karaboga et al. 2024, S. 88).[29]

Es kann davon ausgegangen werden, dass die Technikentwicklung weiter voranschreiten wird. Während die Autor:innen der aktuellen TA-Swiss-Studie Generative KI zur Erzeugung von Videos auf Basis von Texteingaben noch Mitte 2024 in der Zukunft sahen, wurde Ende 2024 mit Sora von OpenAI ein solches Service der breiten Öffentlichkeit zur Verfügung gestellt.[30] Damit fallen einige Barrieren für die Erstellung von Videos weg, wie etwa der Rechenkapazitäten zum Training, die Suche nach geeigneten Trainingsdaten und nicht zuletzt technisches Knowhow (Karaboga et al. 2024, S. 345). Auch die Entwicklung der Modelle der Generativen KI schreitet voran; so dürfte ein vielversprechendes Feld die sog. bilaterale, hybride oder breite (broad) KI sein, also die Kombination von Generativer KI mit anderen KI-Techniken, besonders logik- und wissensbasierter KI, wie es im Rahmen eines FWF-geförderten Clusters of Excellence in Österreich beforscht wird.[31]

2.3 TECHNIK DES ERKENNENS VON KI-GENERIERTEN INHALTEN

Betrachtet man die vielfältigen Folgen und Risiken, die mit Deepfakes und synthetischen Inhalten verbunden sind, so wird deutlich, dass sowohl Transparenz über das Wesen als KI-generierter Inhalt an sich, die Urheberschaft der Erstellung der Inhalte und deren Verbreitung als auch der Schutz der Bürger:innen vor Mis- und Desinformation gewährleistet werden muss. Somit sind sowohl die Authentizität als auch die Herkunft der Inhalte offenzulegen, beispielsweise für die Durchsetzung von Urheberrechten bei der Verwendung von geschützten Daten im Trainingsprozess sowie als Nachweis der Integrität der Inhalte (d. h. Unversehrtheit und Unverfälschtheit), um Manipulationen erkennen zu können. Jüngst verabschiedete Rechtsnormen auf EU-Ebene schreiben derartiger Kennzeichnungen bereits in einigen Fällen vor.

---

29 Siehe beispielsweise die britische Firma Metaphysic, *metaphysic.ai*.

30 *openai.com/index/sora-is-here/*.

31 *vcla.at/2024/05/bilateral-artificial-intelligence/*; siehe auch Abschnitt 6.2.4.

Doch wie kann eine Kennzeichnung kontrolliert werden oder eine Unterlassung nachgewiesen werden? Zur Prüfung dieser Fragestellungen gibt es verschiedene technische Ansätze, um mit einer gewissen Genauigkeit feststellen zu können, ob ein Text, Bild, Audio oder Video von Generativer KI erzeugt wurde oder das Ergebnis KI-gestützter Manipulation ist, etwa indem Artefakte der erzeugenden Methoden erkannt werden (siehe Tabelle 1).

Tabelle 1: Überblick über Ansätze zur Erkennung von KI-generierten Inhalten

Ansatz	Ziel	Herausforderungen	Beispiele
Kennzeichnung von authentischen Inhalten	Ausschließen, dass Inhalte KI-generiert sind	Standardisierung, Robustheit	Content Authenticity Initiative
Kennzeichnung von KI-generierten oder KI-manipulierten Inhalten	Eindeutige Feststellung, dass Inhalte auf KI zurückzuführen sind	Standardisierung, Robustheit	Coalition for Content Provenance and Authenticity, SynthID, AudioSeal
Erkennen von Merkmalen, die wahrscheinlich auf KI zurückführbar sind	Einschätzung, ob nicht gekennzeichnete Inhalte KI-Ursprung haben	Kontinuierliche Weiterentwicklung, Robustheit, Leistungsfähigkeit	Deepware, Reality Defender, SensityAI, DuckDuckGoose

In diesem Abschnitt gehen wir zuerst vertiefend auf die Kennzeichnung von Inhalten – entweder als KI-generiert oder als dezidiert nicht KI-generiert, also authentisch – ein. Danach widmen wir uns aktuellen Ansätzen, KI-generierte oder -manipulierte Inhalte zu erkennen. Dabei werden auch die Herausforderungen berücksichtigt und die Leistungsfähigkeit der Methoden diskutiert.

2.3.1 KENNZEICHNUNG DER HERKUNFT VON INHALTEN

Proaktive bzw. präventive Methoden zur Authentifizierung von echten oder originalen Inhalten dienen der Kennzeichnung und Nachverfolgung der Herkunft (Provenance) der Inhalte. Dabei wird bei oder nach der Erstellung des Inhalts ein Signal eingefügt, das dann von einem Dritten erkannt oder interpretiert werden kann. Bei dieser proaktiven Kenntlichmachung handelt es sich um eine Offenlegung durch Akteur:innen, die an der Entwicklung, Erstellung und Verteilung von Inhalten beteiligt sind. Diese kann aber beispielsweise durch Plattformen zur Information von Nutzer:innen verwendet werden. Man unterscheidet zwischen sichtbaren und unsichtbaren Signalen. Abbildung 3 gibt einen Überblick über Methoden zur Kennzeichnung von Inhalten:

Die Gesetzgeber haben die Möglichkeit zur Überprüfung der Herkunft von Inhalten derzeit in der EU primär durch Offenlegungs- und Kennzeichnungspflichten verankert. Der neue AI-Act[32] regelt dies für Anbieter und Betreiber von KI-Systemen bei der Erstellung derartiger Inhalte. Für Online-Plattformen und deren Nutzer:innen wiederum regelt das Gesetz für Digitale Dienste (Digital Ser-

[32] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, laying down harmonised rules on artificial intelligence and amending Regulations […] (Artificial Intelligence Act), ABl. L vom 12.7.2024, *eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689*.

vices Act, DSA)[33] die Verteilung generierter oder manipulierter Inhalte. So sollen synthetische Inhalte bei der Erstellung in einem maschinenlesbaren Format durch den Anbieter von KI-Systemen gekennzeichnet werden (Artikel 50 (2) AI-Act). Bei der Verbreitung generierter Inhalte durch die Nutzer:innen von sehr großen Online-Plattformen sieht Artikel 35.1 (k) des DSA eine „auffällige Kennzeichnung" vor (vgl. auch Abschnitt 6.1).

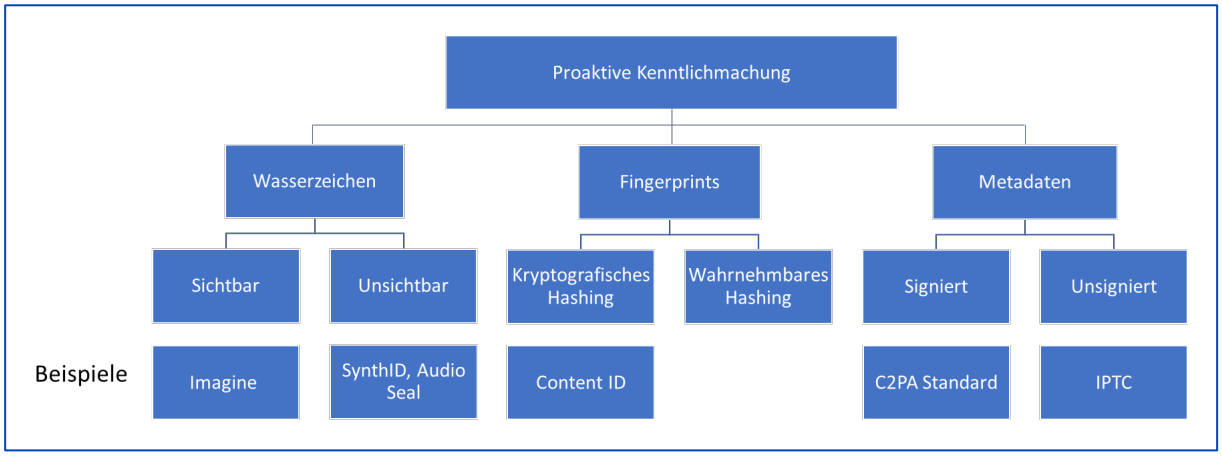


Abbildung 3: Arten und Beispiele für Proaktive Kenntlichmachung
Quelle: abgeändert nach partnershiponai.org[34]

Die Kennzeichnung ist für Menschen direkt oder mit Hilfe von Software erkennbar

Weder im AI-Act noch im Digital Services Act werden spezifische Vorgaben über die detaillierte Technik der Kenntlichmachung gemacht. Man unterscheidet zunächst zwischen einer „auffälligen" und einer „maschinenlesbaren", aber für Menschen unsichtbaren Kennzeichnung. Zu den auffälligen oder auch direkten Kennzeichen zählen neben entsprechenden Markierungen auch kontextuelle Informationen oder Warnhinweise. Im Gegensatz dazu dienen maschinenlesbare oder indirekte Methoden dem technischen Umgang und der Kontrolle von synthetischen Inhalten hinsichtlich der digitalen Authentifizierung, Nachverfolgung, Verteilung und Erkennung. Im Erwägungsgrund 133 des AI-Act werden unterschiedliche Techniken, wie „Wasserzeichen, Metadatenidentifizierungen, kryptografische Methoden zum Nachweis der Herkunft und Authentizität des Inhalts, Protokollierungsmethoden, Fingerabdrücke oder andere Techniken, oder eine Kombination solcher Techniken" genannt.

Nachweis der Originalität, Herkunft und Echtheit durch digitale Signatur

Authentifizierungstechniken wie Wasserzeichen, Fingerprints und Metadaten ermöglichen es, die Originalität oder die Urheberschaft eines Bildinhalts nachzuvollziehen. Allerdings lässt sich so nicht verhindern, dass eine Quelle das Material selbst fälscht, bevor es signiert oder anderweitig authentifiziert wird. Man versucht deshalb, die Echtheit und nicht nur die Originalität von Bild- oder Audioinhalten zu gewährleisten, in dem man die Inhalte z. B. direkt beim Aufnah-

[33] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), ABl. L 277, 27.10.2022, *eur-lex.europa.eu/legal-content/EN/TXT/HTML /?uri=CELEX:32022R2065*.

[34] *partnershiponai.org/resource/glossary-for-synthetic-media-transparency-methods-part-1/*.

meprozess in der Kamera mit einer digitalen Signatur kennzeichnet (BSI 2024). Es besteht jedoch weiterhin die Möglichkeit, dass ein mittels Signatur belegbar echtes Foto auch eine nicht-echte, d. h. mit herkömmlichen Mitteln manipulierte Situation abbilden kann. Außerdem können Signaturprozesse umgangen werden bzw. der digitale Signaturprozess nachkonstruiert werden (Karaboga et al. 2024, S. 89).

Technische Verfahren zur Erzeugung von Metadaten

Es gibt eine Vielfalt an technischen Verfahren zur Erzeugung von Wasserzeichen, Fingerprints und Metadaten mit jeweils Vor- und Nachteilen (Vasse'i/Udoh 2024, S. 27ff). Einige davon werden nun detaillierter vorgestellt.

Bei einer *Metadaten-Kennzeichnung* werden die Informationen über den Urheber, das Datum oder die benutzte Software in den Metadaten und nicht in den Medieninhalten selbst gespeichert, wobei die signierten Metadaten mit einer sicheren Verschlüsselung gespeichert werden. Allerdings ist es möglich, diese Informationen von den Inhalten zu entfernen. Beispielsweise werden solche Metadaten in der Regel gelöscht, sobald ein Bild bei einer Social-Media-Plattform hochgeladen wird.

Wasserzeichen

Für die Erzeugung von Wasserzeichen gibt es unterschiedliche Techniken. Von Bedeutung sind insbesondere *kryptographische Wasserzeichen*, bei denen eine Einschreibung in kryptographische Funktionen des Inhalts erfolgt. Diese Wasserzeichen können dann nur durch einen Verschlüsselungs-/Entschlüsselungsprozess überprüft, entfernt oder geändert werden können. Dieses Verfahren wird insbesondere bei Bild-, Video- und Audio-Inhalten angewendet. Beispiele dafür sind das sogenannte AudioSeal für Audio-Inhalte und KI-generierte Sprache und Googles SynthID.[35] (Vasse'i/Udoh 2024, S. 29).

Fingerprint

In Bezug auf Daten sind sog. Fingerprints das Ergebnis eines anderen kryptografischen Vorgangs: Für eine (große) Datei wird ein eindeutiger (und kleinerer) Wert berechnet. Wie ein Fingerabdruck beim Menschen ist dieser Wert dabei mit sehr hoher Wahrscheinlichkeit nur für diese Datei erzeugbar. Das heißt, wird die Datei modifiziert, ändert sich auch der Fingerprint (auch Hash genannt). Bei synthetischen Inhalten können Fingerprints in einer Datenbank oder auch in einer Blockchain gespeichert werden.[36] In weiterer Folge können von beliebigen Inhalten auf gleiche Weise Fingerprints berechnet werden: Befindet sich der Fingerprint eines Inhalts in der Datenbank, weiß man schließlich, dass es ein KI-generierter Inhalt ist. Im Gegensatz zu Wasserzeichen sind Fingerprints also nicht in die Inhaltsdatei selbst eingebettet. Meistens handelt sich bei den digitalen Kennzeichnungen um Prototypen in der Testphase, um unterschiedliche Kriterien wie Vertrauenswürdigkeit von Inhalten, Rückverfolgbarkeit oder Authentizität von Inhalten zu fokussieren (Aïmeur et al. 2023).

In der Regel müssen Nutzer:innen bei jedem Anbieter einzeln nachprüfen, wie die Inhalte erstellt wurden, weil eine einheitliche und übergreifende Lösung noch fehlt. Für wirksame Authentifizierungstechniken benötigt man deshalb die internationale Zusammenarbeit von Tech-Firmen. Die „Content Authenticity Initiative" (CAI) ist ein von Adobe geführtes Konsortium bestehend aus vielen Zeitungen, Firmen, Kameraherstellern und Bildverwertern (Content Authenticity Initiative 2024).[37]

---

35 *deepmind.google/technologies/synthid/*.
36 *coingeek.com/fusing-blockchain-with-ai-to-fight-bias-and-deepfakes/*.
37 *contentauthenticity.org/our-members*.

Technische Standardisierung

Adobe ist neben Microsoft oder Google auch Mitglied der „Coalition for Content Provenance and Authenticity", welche technische Standards zur Medientransparenz entwickelt. Zu nennen ist hier der C2PA-Standard, der kryptografische Wasserzeichen und die Einbettung von Metadaten kombiniert. Der Standard ermöglicht das Einfügen eines Wasserzeichens in digitale Inhalte, so dass kryptografisch überprüfbare Informationen gespeichert und abgerufen werden können, um die Herkunft und Authentizität der Inhalte zu überprüfen. Open AI wendet beispielsweise diesen Standard für Bilder an, die mit der Schnittstelle von DALL-E 3 oder ChatGPT erstellt wurden.[38] Damit können Nutzer:innen über die Seite „Content Credentials Verify"[39] prüfen, ob ein Bild etwa mit DALL-E 3 generiert wurde. Es ist jedoch

> *„entscheidend, dass offene Standards mit der Beteiligung aller Stakeholder, insbesondere der potenziell Betroffenen, geschaffen werden. Darüber hinaus sollte eine Koalition mächtiger Akteure wie Meta, Google und Adobe nicht zu Torwächtern der Inhaltsintegrität werden."*[40]

Besonderheiten maschinenlesbarer Wasserzeichen

Einer der größten Vorteile maschinenlesbarer Kennzeichnung gegenüber der direkt sichtbaren Kenntlichmachung besteht darin, dass sie schwer oder fast gar nicht zu löschen sind, ohne den Inhalt zu beschädigen, also zu verändern. Wenn sie jedoch für Menschen nicht wahrnehmbar sind, benötigt man spezielle Methoden und Erkennungsprogramme um festzustellen, ob es sich um synthetische Inhalte handelt bzw. welche Provenienz der Inhalt hat.

Anforderungen an Wasserzeichen

Die Kennzeichnung von generierten multimodalen Inhalten aus Bildern, Videos und Texten mit Wasserzeichen ist jedoch voraussetzungsvoll, denn eine Detektion soll auch nach datenverarbeitenden Prozessen beispielsweise eine Datenkompression oder Skalierung sichergestellt werden. Gleichzeitig sollen Wasserzeichen gegenüber bösartigen Angriffen resistent sein. Expert:innen sprechen davon, dass Wasserzeichen „robust" sein müssen (Wang et al. 2023). Es wird auch argumentiert, dass die Entwicklung von robusten Wasserzeichen prinzipiell nicht möglich ist (Zhang et al. 2024).

> *„Generative data is generated by complex deep learning models, and it is a challenge to embed robust watermarks in different data generated by these models without affecting the generation quality. Finally, the added watermarks suffer from multiple attacks. Robust watermarking needs to be resistant to various attacks, e.g., data modification, compression, cropping, and adversarial attacks, and it needs to be ensured that the watermark remains detectable even after the data has been processed."*
> (Wang et al. 2023, S. 14).

Dieses Problem betrifft auch Fingerprints: Da bereits minimale Veränderungen der Datei zu einem anderen Fingerprint führen, bergen schon minimale Veränderungen wie Komprimierung das Risiko, dass sich der Hash verändert und somit vom in der Datenbank gespeicherten Wert abweicht – und somit dieser Ansatz fehlschlägt.

---

38 *help.openai.com/en/articles/8912793-c2pa-in-dall-e-3*.

39 *contentcredentials.org/verify*.

40 Zitat von Molavi Vasse'i in: *foundation.mozilla.org/de/blog/mozilla-research-watermarking-content-labeling-struggle-to-effectively-distinguish-ai-generated-content/*.

2.3.2 GENERIERTE UND MANIPULIERTE INHALTE ERKENNEN

Detektion von nicht-authentischen Inhalten mittels charakteristischer Abweichungen (Artefakte)

Zu den Detektionsmethoden *nach* der Erstellung von generierten oder manipulierten Inhalten zählen manuelle und automatisierte Verfahren, bei denen charakteristische Abweichungen von authentischen Inhalten, sogenannte Artefakte erkannt werden. Es handelt sich dabei u. a. um Unschärfen in Fotos oder Videos insbesondere bei Zähnen oder Augen, Brüchen zwischen Vordergrund und Hintergrund von Gesichtern, inkorrekte Kopfhaltungen oder Augenzwinkern sowie minimale Abweichungen bei der Lippensynchronisation. Diese Inkonsistenzen sind oft für den Menschen mit dem bloßen Auge nicht mehr erkennbar, insbesondere bei der hohen Anzahl an qualitativ hochwertigem audiovisuellem Material in den sozialen Medien in Echtzeit-Situationen. Darüber hinaus ist davon auszugehen, dass es mit fortschreitender Technikentwicklung zu immer weniger Artefakten kommen wird. Eine automatisierte Detektion erfolgt dann unter Zuhilfenahme von KI-Methoden zur Klassifizierung der generierten Inhalte

Eingeschränkte Leistungsfähigkeit der Detektoren

Es muss jedoch berücksichtigt werden, dass beispielsweise komprimierte Daten die Identifizierung von generierten Inhalten erschweren und dass die Leistungsfähigkeit derartiger Detektionssysteme zum Teil weit hinter den Erwartungen bleiben (Groh et al. 2022, S. 118; Karaboga et al. 2024, S. 104f und 113f). Weiterhin handelt es sich bei den Ausgaben der Detektionsmethoden meist um statistische Wahrscheinlichkeiten und keine binären Ja/Nein-Aussagen. Die Nachvollziehbarkeit ist zudem eingeschränkt, da die angewendeten Deep-Learning-Verfahren Black-Box-Methoden darstellen, die auch von Expert:innen nur allgemein erklärbar, aber nicht im Detail nachvollziehbar sind. In der Forschung wird derzeit über neue Möglichkeiten durch erklärbare White-Box-Modelle berichtet (Nguyen et al. 2022, S. 13).

Detektionsmethoden müssen ständig aktualisiert werden

Karaboga et al. (2024, S. 109) listen insbesondere sechs Anbieter von Detektoren für Deepfakes auf: Deepware[41], Reality Defender[42], SensityAI[43], DuckDuckGoose[44], DeepFake-o-meter[45] und MeVer[46]. Auch in Österreich wird aktiv an ähnlichen Werkzeugen gearbeitet, etwa in den Forschungsprojekten defalsif-AI[47] und defame Fakes[48]. Nguyen et al. stellen insgesamt 24 unterschiedliche Techniken zur Detektion von Deepfakes in Bildern oder Videos vor. Die Qualität generierter Bilder und Videos verbessert sich allerdings immer weiter, so dass auch die Leistungsfähigkeit von Detektionsmethoden ständig verbessert werden muss (Nguyen et al. 2022, S. 12).

> *„It is noticeable that a battle between those who use advanced machine learning to create deepfakes with those who make effort to detect deepfakes is growing."* (ibid.)

Bei der Detektion von generierten Inhalten besteht außerdem die Herausforderung, dass es sich um unterschiedliche und komplexe Arten von Daten wie Texte, Bilder, Audios oder Videos handelt. Automatisierte Methoden müssen somit unterschiedliche Techniken und Anwendungen abdecken (Wang et al. 2023,

41 *deepware.ai*.
42 *realitydefender.com*.
43 *sensity.ai*.
44 *duckduckgoose.ai*.
45 *zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/*.
46 *mever.gr*.
47 *science.apa.at/project/defalsifai/*.
48 *deepfakes.at*.

S. 14). Detektionsmethoden beruhen üblicherweise auf einer Mustererkennung, die jedoch gerade bei der Erkennung der von großen Sprachmodellen generierten multimodalen Inhalten erschwert ist:

> *„Large models have higher generative power and creativity, making the generative data more difficult to distinguish."* (Wang et al. 2023, S. 12)

Eine Studie zeigte für die Detektion von generierten Texten mittels ChatGPT, dass alle Texte als von Menschen geschriebene Texte klassifiziert wurden (Wang et al. 2023, S. 9). Darüber hinaus besteht bei generativen Modellen das Problem, die Spuren generierter Daten, sog. Fingerprints, zu identifizieren, da die Modelle iterativ optimiert werden und sich kontinuierlich verändern. Detektionsmethoden müssten deshalb in Echtzeit angepasst werden:

> *„Generative data contains traces (referred to as fingerprints) left by generative models, allowing researchers to detect and attribute the data based on these fingerprints. However, as generative models undergo iterative optimization, these fingerprints also continuously evolve. Therefore, new detection methods must be updated in real time."* (Wang et al. 2023, S. 12)

Tests von Detektoren zum Nachweis von generierten Texten

Weber-Wulff et al. (2023) berichteten weiterhin über viele falsche Ergebnisse bei der Detektion von generierten Texten mittels 14 unterschiedlicher Programme[49] Die Autor:innen zeigten, dass die Systeme sehr einfach zu überlisten sind und dass es letztlich keine Beweise für den Einsatz von KI bei der Texterstellung gibt. Das schließt aber auch aus, dass man sich nicht gegen Vorwürfe zum Einsatz von Textgeneratoren verteidigen kann.

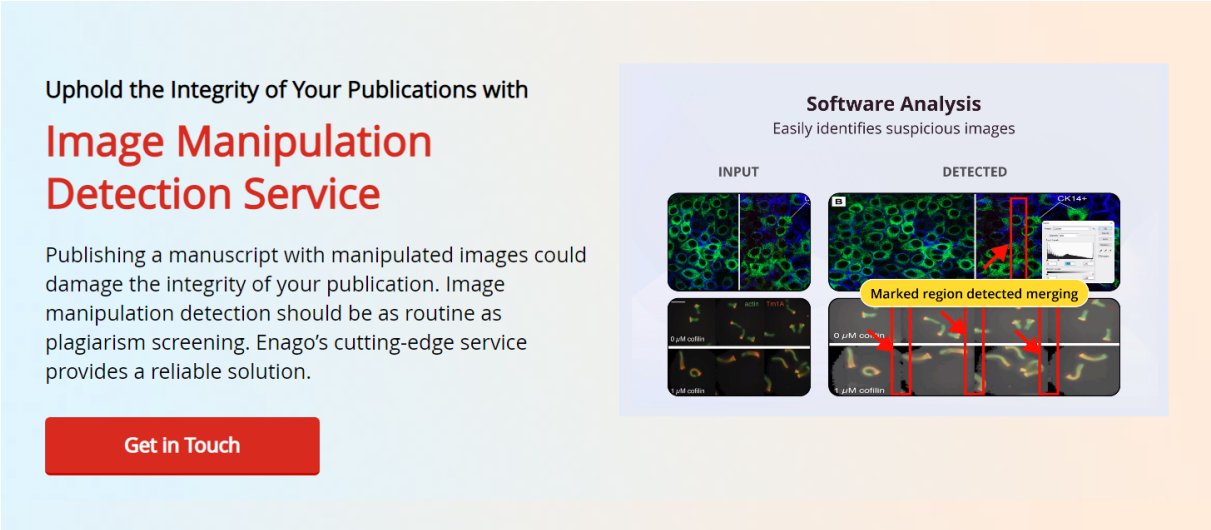


Abbildung 4: Beispiel für den Nachweis einer Bild-Manipulation
Quelle: *enago.com/publication-support-services/image-manipulation-detection*

[49] Check for AI, Compilatio, Content at Scale, Crossplag, DetectGPT, Go Winston, GPT Zero, GPT-2 Outpu Detector Demo, OpenAI Text Classifier, Plagiarism Check, Turnitin, Writeful GPT Detecto, Writer, Zero GPT.

Zwischenfazit

Zusammenfassend kann festgehalten werden, dass sowohl die Authentifizierung vertrauenswürdiger Inhalte als auch die Identifizierung von generierten Inhalten nicht zuverlässig, teilweise technisch überwindbar (Karaboga et al. 2024, S. 105ff) und letztlich nicht effektiv sind. Deshalb sollten diese technischen Verfahren am besten in Kombination mit weiteren rechtlichen oder auch pädagogischen oder organisatorischen Maßnahmen eingesetzt werden (Vasse'i/Udoh 2024; siehe auch Kapitel 6).

3 CHANCEN UND VISIONEN GENERATIVER KI FÜR DIE DEMOKRATIE

Die Digitalisierung von Information und Kommunikation, insbesondere das sich rasant entwickelnde Internet und der Mobilfunk, wurde Anfang des Jahrtausends von vielen als große Chance für demokratische Prozesse betrachtet, beispielsweise im Kontext von Konzepten wie „liquid democracy“[50]. Obwohl seitdem in vielen Fällen Ernüchterung eingetreten ist, widmet sich dieses Kapitel vorrangig den aktuellen positiven Visionen und potenziellen Chancen von Generativer KI für die Demokratie. Erst in den folgenden Kapiteln werden mögliche negative Auswirkungen im Mittelpunkt stehen. Die hier recherchierten und analysierten Aspekten können in vier Kategorien eingeteilt werden, siehe Tabelle 2:

Tabelle 2: Überblick über die Potenziale Generativer KI für die Demokratie

EINSATZ IN DEN (KLASSISCHEN) MEDIEN – Recherchetool – Automatisierter Journalismus – Informationsaufbereitung – Personalisierte News	**DEMOKRATISIERUNG VON INFORMATION** – Sprachübersetzung und -vereinfachung, Barrierefreiheit – Staatsbürgerliche Bildung – Automatisierte Zusammenfassungen – Informationssuche
DISKURSVERBESSERUNG – Verbesserung des politischen Engagements – Konstruktive Diskussionen – Konsenserzielung – Vereinfachte Kontaktaufnahme mit Politiker:innen – Bürger:innen-Beteiligung	**WERKZEUGE FÜR DEN POLITISCHEN ALLTAG** – Automatisierte Protokollerstellung – Dokumentenversionen in einfacher Sprache – Recherchetool – Dokumentenanalyse – Politikberatung – Roboanrufe im Wahlkampf – Microtargeting

3.1 EINSATZ IN DEN (KLASSISCHEN) MEDIEN

Großes Potenzial im Journalismus

Generative KI hat großes Potenzial im Journalismus (vgl. z. B. Pig 2023). KI wird bereits seit Jahren im Nachrichtenbereich eingesetzt, zum Beispiel in Empfehlungssystemen oder bei der Nachrichtenverteilung (Arguedas/Simon 2023). Mithilfe von Generativer KI können heute bereits viele Aufgaben im Journalismus effizienter gestaltet werden, wie z. B. die Nachrichtenbeschaffung oder Transkription, die Beobachtung Sozialer Medien usw. (Gupta et al. 2024). Auch die zentrale Tätigkeit der Journalist:innen, die Recherche, profitiert schon heute durch KI-gestützte Tools. Die Automatisierung dieser Aufgaben schafft potenziell mehr

[50] *de.wikipedia.org/wiki/Liquid_Democracy*.

Raum für anspruchsvollere intellektuelle Tätigkeiten (Baldassarre et al. 2023). Einige Studien argumentieren jedoch, dass KI möglicherweise nicht zwangsläufig dazu führt, dass Nachrichtenmitarbeiter:innen sich mit tiefergehender oder qualitativ hochwertigerer Berichterstattung beschäftigen können. Stattdessen wird die eingesparte Zeit wahrscheinlich schnell durch neue oder zusätzliche Aufgaben in Anspruch genommen (Simon 2024) oder es wird Personal eingespart.

KI-basierte Services zur Recherche und Informationsaufbereitung

Die österreichische Nachrichtenagentur APA entwickelt und nutzt selbst zahlreiche KI-basierte Services und bietet diese ihren Kund:innen in der Medienbranche an,[51] beispielsweise Produktionstools (wie Spracherkennung und Gesichtserkennung), die Verbesserung von Journalismus und Kommunikation durch KI-gesteuerte Lösungen (wie Inhaltsüberwachung und Fake-News-Erkennung) sowie fortlaufende Forschung und Governance durch Studien, Richtlinien und Partnerschaften (APA 2023; vgl. auch Pig 2023). Weiters stellen einige Studien fest, dass sowohl ChatGPT als auch Bard (heute bekannt als Gemini) wahre Content-Erzeugungskraftwerke seien. ChatGPT wird in Journalismus und Marketing eingesetzt, indem es Artikel, Social-Media-Inhalte und sogar Videoskripte produziert. Bard wird vor allem kreativ eingesetzt und dient zur Erschaffung von Gedichten, Skripten und ansprechender Blogbeiträgen (EPTA 2023).

Beispiele für automatisierten Journalismus (Robojournalismus)

In diesem Zusammenhang wird bisweilen auch von Robojournalismus[52] gesprochen, bei dem es um KI-generierte Artikel geht. In der Tat ist Robojournalismus oder automatischer Journalismus mittlerweile in vielen Nachrichtenredaktionen weltweit angekommen. Nachrichtenagenturen wie AP (Associated Press), NTB (Norwegische Nachrichten-Agentur), Ritzau und PA (The Press Association) nutzen automatisierte Systeme zur Erstellung von Texten, z. B. in den Bereichen Bilanzberichterstattung, Wahlen und Sport. Es ist vorstellbar, dass der Anwendungsbereich über die Börsen- und Sportberichterstattung auch das Feld der Berichterstattung über Parlamentsdebatten erreicht. Das britische Projekt RADAR generiert sogar regionalisierte Inhalte auf Basis von Gesundheitsdaten und der Mitteldeutsche Rundfunk (MDR) setzt bei der Wahlberichterstattung teilweise auf Automatisierung (Schell 2022). Grundsätzlich ist die Möglichkeit der KI, die Effizienz in Nachrichtenorganisationen zu steigern, ein zentraler Anreiz für ihre Übernahme (Simon 2024). Der ORF hat für sein selbst entwickeltes multifunktionales KI-Tool AiDitor kürzlich sogar einen internationalen Preis gewonnen. Mit dem AiDitor „können auf Basis von Links oder recherchierten Inhalten weitere Formate wie Onlineartikel, TV- und Radiotexte, Teletextbeiträge, Social-Media-

---

51 *apa.at/produkt/companygpt-by-apa/*.

52 Siehe *parlament.gv.at/dokument/fachinfos/zukunftsthemen/006_robojournalismus.pdf*. Zum Begriff des Robojournalismus: Das aufkommende Forschungsfeld der Automatisierung in der journalistischen Textproduktion hat sich noch nicht auf einen endgültigen Begriff geeinigt. Verschiedene Bezeichnungen wie „computergestützter Journalismus“, „algorithmischer Journalismus“, „maschinengetriebener Journalismus“ und „automatisierter Journalismus“ wurden vorgeschlagen. Zusätzlich wurden Begriffe wie „Robo-Writing“ und „robotischer Reporter“ verwendet. Die APA hat sich entschieden, das Wort „Roboterjournalismus" abzulehnen und stattdessen „Automatisierter Journalismus“ zu verwenden, da dies besser repräsentiert, was der Begriff bedeutet, nämlich eine Variante des Datenjournalismus mit einem hohen Automatisierungsgrad. Damit wird auch die suggestive Bedeutung eines „autonomen maschinellen Akteurs“ vermieden. Anstatt vorhandene Inhalte zur Generierung neuer Inhalte zu verwenden, bezieht sich der automatisierte Journalismus, wie von der APA verwendet, auf die Erzeugung von originären Inhalten aus organisierten Daten (Schell 2022).

Postings und Überschriften generiert werden. Weiters ist es mit dem Tool möglich, Texte zu transkribieren und die Audioqualität von Aufnahmen zu verbessern. Der AiDitor erkennt außerdem Zitate und kann in mehr als 40 Sprachen übersetzen."[53]

Die Rolle menschlicher Journalist:innen

Präzise und zuverlässige Daten sind für eine qualitative journalistische Arbeit unerlässlich. Daher ist das Vertrauen in Werkzeuge wie ChatGPT, die Inhalte aus einer nicht verifizierten Datenbank erstellen, nur bedingt für guten Journalismus geeignet (Simon 2024). Dennoch kann der Journalismus Generative KI im weiteren Sinne nutzen, indem spezifische und angepasste Tools verwendet werden, die intern trainiert sind, um kontrollierte Ergebnisse zu erzeugen. Diese müssen selbstverständlich von menschlichen Journalist:innen gründlich überprüft werden. Laut APA ist es noch nicht möglich, durch die gängigen generativen Werkzeuge und Maschinen allein, ohne dass Menschen spezifische Regeln vorgeben, faktenbasierte und zuverlässige Informationen zu erzeugen (APA 2023). Effizienzsteigerungen variieren je nach Aufgabe und Kontext. Diese Fortschritte können durch Faktoren wie die Unzuverlässigkeit der KI-Ergebnisse, das Risiko von Rufschädigung durch ungenaue KI-Ergebnisse und die Herausforderungen bei der Automatisierung spezifischer Aufgaben begrenzt sein. Da Nachrichten erfordern, dass Journalist:innen sich schnell anpassen, erschwert dies eine vollständige Automatisierung von Nachrichten durch KI (Simon 2024).

Personalisierte News

Ein weiterer Anwendungsbereich für Generative KI ist die Option, die Medienkonsument:innen durch personalisierte, relevante, zugängliche und ansprechende Inhalte effektiver zu informieren (Arguedas/Simon 2023). So könnte das Phänomen der sog. „Nachrichtenvermeidung" (oder „-enthaltung") verhindert werden, das offenbar durch Nachrichten verursacht wird, die dem Publikum zu deprimierend oder langweilig sind. Forschungsergebnisse zeigen, dass Menschen, die Nachrichten vermeiden, eher an positiver oder lösungsorientierter Berichterstattung interessiert sind und weniger an den großen Geschichten des Tages (Newman 2023). Beispielsweise nutzt die bereits 2017 gegründete Agentur RADAR AI Generative KI, um aktuelle Nachrichten auszuwählen und sie an das Publikum anzupassen, indem die Geschichten in einen breiteren Kontext gestellt werden (PA Media 2022). Auf diese Weise fühlen sich die Menschen mehr mit den Geschichten verbunden und es entsteht ein besseres Verständnis für die globale Situation, anstatt isolierte Nachrichten zu erhalten. Im Zusammenhang mit politischen Angelegenheiten könnte dieses Werkzeug mehr Engagement für aktuelle politische Themen schaffen, für bevorstehende Wahlen sensibilisieren oder politische Meinungen verständlicher machen (Arguedas/Simon 2023).

Kontextualisierte News

Mögliche Risiken des Robojournalismus

Verzerrungen und Machtungleichgewichte?

Diese Werkzeuge sind freilich nicht unumstritten. Medienwissenschaftler:innen warnen vor Verzerrungen und Machtungleichgewichten bei der Nutzung der verwendeten Daten. Lin und Lewis argumentieren etwa, dass KI zwar einen positiven Einfluss auf den Journalismus in drei seiner Hauptbereiche haben kann (Informationsbeschaffung, Auswahl und Produktion sowie Verteilung und Konsum), jedoch nur, solange die folgenden Werte beachtet werden: Genauigkeit, Zugänglichkeit, Vielfalt, Relevanz und Aktualität. Andernfalls könne dies negative Auswirkungen auf die demokratische Teilhabe und Diskussion haben (Lin/Lewis 2022).

[53] *orf.at/stories/3360846/*.

Beschränkung der Meinungsvielfalt und Bias

Nachteile des automatisierten Journalismus wurden noch keiner systematischen TA-Analyse unterworfen, was auch hier nicht geleistet werden kann. Es erscheint freilich möglich, dass durch den Einsatz von KI-Systemen anstelle von Journalist:innen für die Produktion oder Verbreitung von Nachrichten die Vielfalt der Meinungen weiter beeinträchtigt werden könnte. Der durch KI verursachte Bias (vgl. Abschnitt 4.1) könnte die parlamentarische Arbeit verzerren, da Politiker:innen i. d. R. intensive Medienkonsument:innen sind. Diese Modelle werden mit großen Datensätzen trainiert, die Verzerrungen enthalten und daher Minderheits- oder abweichende Standpunkte in politischen Diskussionen übersehen können (EPTA 2023). Während KI die Nachrichtenorganisationen transformiert, wird sie auch die Öffentlichkeit beeinflussen, die für die Demokratie entscheidend ist und für die die Nachrichtenorganisationen als Gatekeeper fungieren. Diese Entwicklung wird durch Entscheidungen von zwei Gruppen beeinflusst: diejenigen, die direkt die Nachrichten kontrollieren (Führungskräfte, Manager:innen, Journalist:innen), und immer mehr die, die eine indirekte oder keine Kontrolle auf die Arbeit rund um Nachrichten haben (Technologieunternehmen, Regulierungsbehörden und die Öffentlichkeit) (Simon 2024).

3.2 DEMOKRATISIERUNG VON INFORMATION

Information und (darauf aufbauendes individuelles oder kollektives) Wissen sind essentielle Bestandteile jedes demokratischen Diskurses, jeder Meinungsbildung und jeder Entscheidungsfindung. Digitale Hilfsmittel wie Datenbanken oder Expertensysteme und seit geraumer Zeit das Internet mit seinen Suchmaschinen spielen dementsprechend schon lange eine zentrale Rolle. KI-Systeme und insbesondere Anwendungen Generativer KI haben nun das Potenzial, auf diesem Feld zukünftig einen qualitativen Beitrag zu leisten. Schon im Zusammenhang mit dem (frühen) Internet wurde von „Demokratisierung" gesprochen (Thiel 2020), womit gemeint ist, dass der Zugang und die Verfügbarkeit von Information aller Art von wenigen Privilegierten auf potenziell alle Bürger:innen ohne Schranken ausgedehnt wird. Generative KI ist potenziell in der Lage, diese Zugänglichkeit deutlich zu verbessern.

Informationsgewinnung

Generative KI bietet niedrigschwellige Tools zur Informationsgewinnung, die potenziell weit über die bisherigen Suchmaschinen hinausgehen. Vor allem die Suche in natürlicher Sprache bzw. über Chatbots stellt für den Alltagsgebrauch einen Fortschritt dar.

Sprachübersetzung und -vereinfachung

Eine weitere typische Anwendung Generativer KI sind Sprachübersetzungen. Damit wird es für alle viel leichter, Informationen aus dem Internet in der eigenen Sprache zu konsumieren, was gerade in einer multilingualen Gesellschaft wie der Europäischen Union sehr nützlich sein kann und auch die interkulturelle Kommunikation erleichtert (Baldassarre et al. 2023). Diese neuen Systeme sind darüber hinaus in der Lage, Fachdokumente in einfachere[54] bzw. Nutzer:innen-adäquate Sprache zu übertragen, sogar vorzulesen und an kulturelle Hintergründe und Kontexte zielgruppenspezifisch anzupassen. Gut geeignet dürfte Generative KI auch zur Herstellung von Barrierefreiheit sein; weiters wird die Umsetzung in Bilder und Videos möglich. Damit werden potenziell

Barrierefreiheit

[54] Das österreichische Parlament nutzt zur Vereinfachung seiner Webinhalte den Dienst *capito.eu*, siehe *parlament.gv.at/services/barrierefreiheit/einfache-sprache/*.

breitere Gesellschaftsschichten erreichbar – was angesichts der zentralen Eigenschaft einer Demokratie wesentlich ist, sind doch alle Bürger:innen unabhängig vom formalen Bildungsniveau ab einer gewissen Altersgrenze aufgerufen, sich mit ihrer Stimme am demokratischen Willensbildungsprozess zu beteiligen. Somit könnte der Einsatz Generativer KI potenziell zu mehr Inklusion beitragen.

Automatisierte Zusammenfassungen

Auch die Erzeugung von Zusammenfassungen langer (Fach-)Texte, die ansonsten aus Zeitmangel kaum gelesen würden, kann dazu beitragen, das dort enthaltene Wissen im demokratischen Diskurs wirksam werden zu lassen. Es besteht also das Potenzial, dass Generative KI den Zugang zu Ideen und Wissen beschleunigt, indem sie eine Vielzahl von Inhalten schnell aggregiert und den Suchprozess vereinfacht, was den Menschen helfen kann, effizienter neue Informationen zu sammeln (STAA 2023). Beispielsweise gibt es seit kurzem die Plattform parlament.fyi, die nach Angaben des dahinterstehenden „Vereins zur Förderung von digitaler Politik und Künstlicher Intelligenz" Österreichs erste KI-basierte Plattform über Politik sei: KI „fasst Beschlüsse und Reden aus dem Nationalrat einfach zusammen und liefert kompakte Analysen zu Sprachmustern und Argumenten der Parteien. Ergebnisse aus der Politik werden damit für alle leichter zugänglich und transparent." Noch steckt die Plattform in den Kinderschuhen, sie hat freilich Potenzial.[55]

Kommunikationserleichterung?

Bürger:innen sind im Idealfall aber nicht nur passive Rezipient:innen von Information, sondern beteiligen sich auch aktiv am demokratischen Diskurs. Hierbei kann Generative KI ebenfalls eine positive Rolle spielen, indem öffentliche Äußerungen aller Art, z. B. Leser:innenbriefe oder die Kommunikation mit politischen Amtsträger:innen mit ihrer Hilfe erleichtert und verbessert werden. Auch auf Seiten der Politik (siehe der folgende Abschnitt 3.3) können Chat- oder Sprachbots die Kommunikation der einzelnen Politiker:innen mit einer Vielzahl von Bürger:innen erleichtern und für beide Seiten wegen der individuellen Ansprache und der Zweiseitigkeit befriedigender machen als Standardbriefe. Dies alles könnte sowohl zu quantitativ als auch zu qualitativ besserem politischem Engagement führen. Freilich ist diese Verbesserung zweischneidig, denn es wird tatsächlich persönliche Kommunikation durch maschinelle ersetzt und der persönliche Kontakt nur mehr vorgetäuscht. Dies könnte letztlich sogar zu einer weiteren Entfremdung vom politischen Diskurs führen.

Staatsbürgerliche Bildung

Insgesamt könnten diese Potenziale von Generativer KI bewirken, dass die Bevölkerung besser informiert ist. Coeckelbergh schlägt etwa ein ChatGPT-artiges Programm zur Unterstützung der staatsbürgerlichen Bildung und zur Zusammenfassung des Wissensstands vor, um damit „zur epistemischen Stärke dieser Säule der Demokratie beizutragen" (2024, S. 87f, Übersetzung durch die Autor:innen).

Bekämpfung von Fehlinformationen

Ob die Methoden der Generativen KI auch selbst dazu beitragen, Des- und Misinformation zu bekämpfen (Coeckelbergh 2024) oder vielmehr selbst weiter zu Fehlinformation beitragen, ist Gegenstand der Abschnitte 2.3 und 6.3.

[55] *parlament.fyi*.

3.3 VERBESSERUNG DES POLITISCHEN DISKURSES

Wie bereits weiter oben erwähnt, kann Generative KI dazu beitragen, ein besseres Verständnis für verschiedene Themen und politische Angelegenheiten zu schaffen. Sie kann auch Informationen über das Spektrum der politischen Ansichten zu einem Thema bereitstellen und damit zu Diversität und Pluralität beitragen (Coeckelbergh 2024). Dieser potenziell positive Effekt gilt sowohl für die Bürger:innen als auch für Politiker:innen, die sehr viele unterschiedliche Agenden mit wenig Zeitressourcen zu bewältigen haben. Darüber hinaus könnte Generative KI dazu beitragen, den demokratischen Diskurs insgesamt zu verbessern. Es sei auch an dieser Stelle betont, dass es aus heutiger Sicht aufgrund der enormen Dynamik der technischen, aber auch der gesellschaftlichen und politischen Entwicklung sowie der Interdependenzen der verschiedenen Strömungen unmöglich ist vorherzusagen, ob dieses Potenzial tatsächlich je realisiert werden kann und wie sich die Risiken durch eine ebenso mögliche Diskursverzerrung auswirken (Abschnitte 4.1 und 4.2).

Generative KI und Bürger:innen-beteiligung

Wir beobachten ein wachsendes Interesse an der Integration von Generativer KI in die Bürger:innenbeteiligung und den offenen politischen Diskurs. Soziale Medien haben unbestritten hohe Relevanz im politischen Diskurs; jedoch fehlt ihnen sicherlich die Absicht und die deliberative Qualität, die für konkrete Problemlösung notwendig sind (Tsai et al. 2024). Stellvertretend für viele – und ähnlich wie vor zwei Jahrzehnten unter der Überschrift „Liquid Democracy", sieht beispielsweise Coeckelbergh (2024) Chancen für KI zur Erneuerung demokratischer Prozesse, insbesondere bei der Unterstützung von Verfahren zur Beteiligung von Bürger:innen.

In Österreich sind bereits viele digitale Werkzeuge und Umfrageplattformen für Partizipation im Einsatz, wie *Mein Parlament*, *oesterreich.gv.at*, *Digitales Amt* usw. Weltweit nutzen bestimmte deliberative Plattformen Generative KI, die über herkömmliche Umfrageplattformen hinausgehen, indem sie aktiv vielfältige Standpunkte zu einer Frage einholen, wichtige Kommentare für die Überprüfung durch die Teilnehmer:innen hervorheben und manchmal kollektive Entscheidungsfindung erleichtern.

Höhere Kooperations- und Beteiligungsniveaus

Die Einbindung der Bürger:innen wurde mit erhöhtem Vertrauen in die Regierung und mehr Kooperation und Beteiligung der Bürger:innen in Verbindung gebracht (O'Brien/Tyler 2020). Forschungsergebnisse des MIT zeigen ebenfalls, dass die Ansprache von Bürger:innen mit höherem Vertrauen in die Regierung und höheren Kooperations- und Beteiligungsniveaus verbunden ist (Tsai et al. 2020). Neue Online-Deliberationsplattformen behaupten, diese Ziele schneller und in größerem Umfang mit reduziertem menschlichem Bias und niedrigeren Kosten zu erreichen. Die erste Generation dieser Plattformen umfasst Funktionen wie offene Kommentare, Upvoting und einfache Umfragen sowie die Organisation von Versammlungen und partizipativem Budgetieren. Einige Plattformen verbessern dies durch den Einsatz von Maschinellem Lernen zur Visualisierung öffentlicher Meinungen, Hervorhebung von Übereinstimmungen und Meinungsverschiedenheiten und Identifizierung struktureller Aspekte zur Förderung von Kompromissen oder Konsens (Tsai et al. 2024).

Beispiel Pol.is (genutzt vom Klimarat)

Ein Beispiel für die Nutzung von KI in der Deliberation ist die Plattform *Pol.is*, die 2022 in Österreich vom Klimarat genutzt wurde, um unter mehr als 5.000 Menschen einen Konsens zu Klimafragen zu finden. Pol.is wurde vom Computational Democracy Project (2024a) entwickelt und ist bekannt für den Einsatz fortschrittlicher computerbasierter Techniken, einschließlich KI, um Echtzeit-Feedback und Diskursanalysen in großem Maßstab zu erleichtern. Diese Plattform verwendet maschinelle Lernmodelle, um die Muster von Zustimmung und Widerspruch unter den Teilnehmer:innen zu interpretieren und zu visualisieren (The Computational Democracy Project 2024b). Pol.is kann groß angelegte Gespräche analysieren und visualisieren: Damit erleichterte es die Einbeziehung einer breiten Meinungsvielfalt und half dem Klimarat, öffentliches Feedback in umsetzbare Empfehlungen für das österreichische Parlament zu integrieren.

Weitere Plattformen und Projekte, die KI (nicht unbedingt Generative KI) verwenden

Es gibt noch weitere Plattformen, die dasselbe Ziel der konstruktiven politischen Deliberation mit Hilfe von KI verfolgen:

- *Assembl*, entwickelt von der französischen Firma Bluenove, nutzt Generative KI zur Zusammenfassung von Diskussionen und zur Identifizierung von Schlüsselelementen aus großen Textmengen (Bluenove 2024).
- Die *Stanford Online Deliberation Platform*, entwickelt von den Teams für deliberative Demokratie und Crowdsourced Democracy an der Stanford University, setzt KI ein, um den Deliberationsprozess zu verbessern, indem sie Beiträge der Teilnehmer:innen moderiert, analysiert und zentrale Punkte zusammenfasst (Stanford Deliberative Democracy Lab 2022).
- *Wikum* wurde von MIT-Forscher:innen entwickelt und verwendet KI, um große Textmengen aus Online-Diskussionen zusammenzufassen und zu organisieren, sodass Benutzer:innen schnell die Hauptpunkte und Trends in der Diskussion erfassen können (MIT CSAIL 2017).
- *ConvoWizard* ist bzw. war ein im Rahmen eines Forschungsprojekts der Cornell University entwickeltes Tool, das zu konstruktiveren politischen Diskussionen beitragen wollte.[56]
- Im Rahmen des netidee SCIENCE Förderprogramms untersucht derzeit ein Team an der WU Wien, wie bei digitalen Gruppenentscheidungen Minderheiten mehr Repräsentation erlangen können und Polarisierung zurückgedrängt werden kann – um in weiterer Folge den politischen Diskurs zu verbessern.[57]

Mögliche Einsatzbereiche

Der Einsatz dieser Arten von Plattformen in der österreichischen Politik könnte unter Bürger:innen vor oder im Vorfeld eines Volksbegehrens genutzt werden, um wichtige Themen zu debattieren und herauszufinden, welche Anliegen die Aufmerksamkeit des Parlaments erfordern. Die Deliberation könnte auch einer Volksbefragung vorausgehen, um den Kontext der Abstimmung und ihre Konsequenzen besser zu kommunizieren. Auch für jede Art von Wahl (lokal, regional oder national) könnte ein Raum für die Deliberation der Bürger:innen im Vorfeld angeboten werden. Risiken bestehen jedoch darin, dass KI-generierte Konsensstatements eher aufgrund ihres freundlicheren oder autoritativeren Tons zu Übereinstimmung führen könnten, als deshalb, weil die Menschen tatsächlich mit dem Inhalt übereinstimmen (Tsai et al. 2024).

---

56 *zissou.infosci.cornell.edu/convokit/convowizard/*.

57 *fwf.ac.at/aktuelles/detail/jan-felix-maly-erhaelt-netidee-science-foerderung-2024*.

Individuelle Kommunikation zwischen Politiker:innen und Bürger:innen (z. B. „Robocalls“)

Der politische Diskurs wird auch durch die kommunikativen Beziehungen zwischen den politisch Aktiven und der Bevölkerung geprägt. Generativer KI, insbesondere Chat- oder Sprachbots, verändern die Kontaktaufnahme mit politischen Repräsentant:innen durch Bürger:innen.[58] Das hat das Potential, die Kommunikation zu verbessern, ist aber, etwa in Form von „Roboanrufen“, nicht unumstritten.[59] Ein aktuelles Beispiel wurde im laufenden US-Wahlkampf bekannt: Der Audio-Bot *Ashley* kann individuelle Telefonanrufe mit potenziellen Wähler:innen durchführen.[60] Im Vergleich zu Microtargeting (siehe Abschnitte 3.4 und 4.3) ist hier Kommunikation in beide Richtungen möglich. Das Potenzial von Generativer KI für die direkte Kommunikation mit Politiker:innen könnte freilich limitiert sein: Wenn Interessengruppen KI nutzen, um Massenkampagnen zu optimieren, indem sie Bürger:innen helfen, E-Mails an Politiker:innen zu personalisieren, könnte es sein, dass die Adressat:innen skeptisch werden, wenn sie vermuten, dass sie mit einer KI kommunizieren. Zusätzlich besteht bei diesen Technologien heute immer noch das Risiko von „Halluzinationen“ und der Bereitstellung von verzerrenden (biased) Informationen (siehe Abschnitt 4.1).

3.4 WERKZEUGE FÜR DEN POLITISCHEN ALLTAG

Schließlich kann Generative KI auch dazu eingesetzt werden, den Politikbetrieb selbst zu unterstützen. Es eröffnen sich insbesondere für den Parlamentsbetrieb, für Parteien und Politiker:innen sowie in der Bürokratie neue Möglichkeiten.

Sitzungsprotokolle

Aufgrund der bemerkenswerten Entwicklung im Bereich der Text-to-Speech- bzw. Speech-to-Text-Technologie werden vermutlich in Zukunft Sitzungsprotokolle praktisch ohne Zeitverlust automatisiert erstellt werden können. Während dafür bislang Stenograph:innen engagiert werden mussten, ist es vorstellbar, dass die neuen Tools in Zukunft bei Bedarf auch in kleineren, ad-hoc einberufenen oder anfangs informellen Sitzungen eingesetzt werden können.

So wurde beispielsweise im März 2024 bei der Vernetzungsveranstaltung „KI in der Hochschullehre“ des Wissenschaftsministeriums Baden-Württemberg, ein KI-Verfahren mit *ChatGPT* zur Dokumentation und Analyse der Veranstaltung eingesetzt. Das experimentelle Verfahren wurde vom Karlsruher Institut für Technologie (KIT) in Zusammenarbeit mit dem Hochschulnetzwerk Digitalisierung der Lehre Baden-Württemberg begleitet (HND BW 2024). Als Ergebnisse dieses Experiments konnte festgehalten werden: Die Qualität der Prompts spielten eine entscheidende Rolle für das Ergebnis. Aber auch die Metadaten und der Kontext waren von großer Bedeutung. Das Verfahren lässt sich schnell durchführen und war kostengünstig. Allerdings hat sich die verwendete KI bei der Generierung wörtlicher Zitate aus den analysierten Vorträgen als äußerst unzuverlässig erwiesen, weshalb unbedingt ein menschlicher Überprüfungsschritt im Prozess integriert werden musste.

---

58 Z. B. *journalofdemocracy.org/articles/how-ai-threatens-democracy/*.

59 *reuters.com/world/us/pastors-secret-codes-us-election-officials-wage-low-tech-battle-against-ai-2024-10-31/*.

60 *reuters.com/technology/meet-ashley-worlds-first-ai-powered-political-campaign-caller-2023-12-12/*.

Zusammenfassung und Aufbereitung von Texten

Da der Arbeitsalltags politisch Aktiver auch durch das Lesen bzw. Scannen unzähliger kürzerer und längerer Schriftstücke, inkl. von Wortprotokollen, geprägt ist, besteht ein Potenzial des Einsatzes Generativer KI in der Zusammenfassung und im Verfassen von leichtfasslicheren Dokumentenversionen. Doch auch hier gibt es bislang noch viele offene Probleme großer Sprachmodelle, etwa die Tendenz, Informationen in der Mitte eines Dokuments zu ignorieren.[61] Auch automatisierte Diskursanalysen über Sitzungen oder die Umformulierung von Medienberichterstattung wären für Politiker:innen und interessierte Öffentlichkeit mitunter ein spannendes Anwendungsfeld für Generative KI.

Recherchetool

Schließlich kann Generative KI auch von Politiker:innen und ihren Beratungsstäben (etwa den parlamentarischen Mitarbeiter:innen der Abgeordneten, dem wissenschaftlichen Dienst des Parlaments oder externen Beratungseinrichtungen) in Zukunft als neues, effizientes Recherchetool bzw. zur raschen Dokumentenanalyse eingesetzt werden. Allerdings sei hier noch vor vorschnellem und unreflektiertem Einsatz gewarnt, da die Tools einer dynamischen Entwicklung unterliegen. Es ist zwar möglich, dass die Ergebnisqualität (also nicht nur die Sprache, sondern vor allem die Validität der Recherche – Stichwort „Halluzinationen“[62]) sukzessive verbessert werden wird, dass dies aber zum heutigen Zeitpunkt keineswegs als gesichert angenommen werden sollte.

Diskursunterstützung, Konsensfindung

Entscheidungsfindung ist ein essentielles Element politischer Tätigkeit. Die im vorigen Abschnitt genannten Tools und Plattformen zur Darstellung des Wissens- und Positionsspektrums sowie zur Unterstützung des Diskussionsprozesses könnten potenziell auch in Parlamenten zum Einsatz kommen. Es wird auch daran gearbeitet, Tools zur Unterstützung der Konsensfindung auf KI-Basis bereitzustellen.

Beispiele

Derzeit gibt es einige Webplattformen in den USA wie *Ballotpedia* oder *Vote Smart*, die den Zugang zu wichtigen Informationen für Wähler:innen verbessert haben. Da diese Websites jedoch schwer zu navigieren sind, versuchen KI-Technologien diese Dienste zu verbessern. Politiker:innen kann KI dabei helfen, Kommentare von Bürger:innen aus öffentlichen Konsultationen oder E-Mails zusammenzufassen. KI-Systeme können Feedback nach verschiedenen Kriterien klassifizieren und so ein besseres Verständnis der öffentlichen Meinung ermöglichen, insbesondere in Kombination mit menschlicher Analyse.[63] Das oben (Abschnitt 3.3) erwähnte *Pol.is* ermöglicht es den Teilnehmer:innen, Meinungen zu bestimmten Themen zu äußern, und anderen, diesen Meinungen zuzustimmen oder zu widersprechen. Das KI-basierte Tool wählt dann Gruppen von Personen aus und stellt diese dar, die einen Konsens finden könnten, anstatt Meinungsverschiedenheiten hervorzuheben.[64]

Microtargeting

Abgesehen vom Potenzial für die Kommunikation durch Chatbots und ähnliche Tools mit den Wähler:innen (siehe oben) gibt es verschiedene und seit einigen Jahren bereits genutzte Tools der politischen Kommunikation insbesondere

---

61 *heise.de/news/Neuer-Ansatz-gegen-Lost-in-the-middle-Problem-von-Sprachmodellen-9701887.html*.

62 Damit ist gemeint, dass etwa ChatGPT bislang auch Quellen und Tatsachen erfindet, wenn sie sprachlich in den Kontext passen (siehe Abschnitt 4.1.2).

63 Siehe *theconversation.com/generative-ai-like-chatgpt-could-help-boost-democracy-if-it-overcomes-key-hurdles-212664* und *europarl.europa.eu/RegData/etudes/BRIE/2023/751478/EPRS_BRI(2023)751478_EN.pdf*.

64 *theguardian.com/world/2020/sep/27/taiwan-civic-hackers-polis-consensus-social-media-platform*.

in Wahlkämpfen. Insbesondere Microtargeting, also die individualisierte Verbreitung von Botschaften an kleine und kleinste Gruppen von Wähler:innen, ist freilich umstritten und wird daher hier unter „Risiken" analysiert (siehe Abschnitt 4.3). In diesem Abschnitt werden wir doch versuchen einige Potenziale des Microtargeting für die Demokratie vorzustellen.

Dieses Phänomen (Microtargeting) entstand durch die allgemeine Nutzung sozialer Netzwerke. Es gibt mehrere potenzielle positive Auswirkungen von Microtargeting auf die Demokratie, insbesondere durch den Einsatz von generativer KI. Generative KI ermöglicht es, politische Botschaften effizienter und zielgerichteter zu verbreiten. Dadurch können Politiker mehr Menschen mit weniger Aufwand erreichen, besonders im Vergleich zu traditionellen Kampagnen wie Straßenwahlkämpfen (Matz et al. 2024). Die Zielgruppen werden anhand von Variablen wie Interessen, Alter, Geschlecht oder Standort gezielt angesprochen. KI kann auch Botschaften individuell anpassen und spezifisch auf die Bedürfnisse und Interessen einzelner Wähler zuschneiden (Zarouali et al. 2022). Dies könnte dazu führen, dass politische Informationen relevanter und ansprechender für die jeweiligen Empfänger sind (Matz et al. 2024). Die Erstellung von Botschaften in verschiedenen Formaten (Texte, Videos, Fotos) kann auch durch Generative KI automatisiert und in großem Umfang personalisiert werden. Die Verwendung von den neuesten Technologien könnte grundsätzlich zu innovativeren und effektiveren Wahlkampagnen führen. Diese neuen Potenziale für die Politik sind noch so unerforscht, dass unklar bleibt, ob und wie sie vollständig umgesetzt werden. Die tatsächlichen Auswirkungen der Kombination von Personalisierung und Emotionalisierung auf politische Diskurse und Meinungsbildung sind weiterhin unklar und Gegenstand laufender Forschung (Dumbrava 2021; Karmasin et al. 2024). (Für die tiefere Analyse, siehe Abschnitt 4.3.)

Data-driven Campaigning

In ähnlicher Weise wird das Phänomen des „Data-Driven Campaigning" oder „Datengesteuerte Kampagnen" als der Aufstieg zunehmend ausgeklügelter, hochgradig zielgerichteter Datennutzungen wahrgenommen, die eingesetzt werden, um die Popularität eines/r Kandidaten/in zu steigern. Diese Praxis wird als transformierend für die Wahlkampfmethoden angesehen und wirft Bedenken hinsichtlich demokratischer Prozesse auf. Aber Wissenschaftler:innen wie Dommett et al. (2023) betonen, dass das Verständnis dessen, was datengetriebener Wahlkampf wirklich ist, nur teilweise vorhanden ist und schlagen eine differenziertere Sichtweise vor. Datengetriebener Wahlkampf tritt in verschiedenen Formen auf, die nicht von Natur aus problematisch sind, angesichts der Vielfalt der Parteien und Länder, die ihn nutzen. Hierbei ist es wichtig, zwischen systemischen, regulatorischen und parteiinternen Variablen zu unterscheiden. In ihrer Studie verwenden sie Belege aus fünf Ländern und entdecken, dass entgegen der weit verbreiteten Meinung die meisten Parteien weder fortgeschrittenes Microtargeting noch komplexe Tests durchführen. Obwohl Daten für moderne Wahlkämpfe von entscheidender Bedeutung sind, ist ihre Nutzung komplexer und kontextabhängiger, als allgemein angenommen wird, birgt aber viele Potenziale zur Effizienzsteigerung im politischen Bereich (Darius/Römmele 2023; Dommett et al. 2023) (siehe Abschnitt 4.3 für eine tiefere Analyse und potenzielle Nachteile von data-driven campaigning).

Wahlkampf-App

KI-Systeme haben Potenzial für Wahlen, indem sie Prozesse wie Wahlempfehlungs-Apps, die Organisation von Wahlen und Prognosen vereinfachen und gleichzeitig die Effizienz der Informationsverbreitung und Wählermobilisierung verbessern (Heesen et al. 2021). Ein Beispiel für den Einsatz von Daten zur Un-

terstützung von Wahlkämpfen, obwohl keine Generative KI verwendet wird, ist die App „CDU connect“, die von der deutschen Partei CDU entwickelt wurde. Sie wurde erstmals für die Bundestagswahl 2017 eingeführt, um den Haustürwahlkampf zu unterstützen (Hurz 2021). Die App nutzt soziologische, statistische und Geodaten, um Wahlkampfaktionen zu optimieren. Registrierte Nutzer:innen können an einem kompetitiven „Spiel“ teilnehmen, bei dem sie Punkte für Wahlkampfaktivitäten wie Hausbesuche oder Social-Media-Posts sammeln (Burger/Jaeger 2017).[65]

3.5 AUSBLICK

Zusammenfassend kann somit festgehalten werden, dass KI-Anwendungen, insbesondere auch Generative KI, durchaus vielversprechendes Potenzial für die Politik und Demokratie haben. Allerdings müssten diese Chancen auch aktiv genutzt und sinnvoll gestaltet werden, um tatsächlich demokratiefreundliche Ergebnisse zu bewirken. Als „Selbstläufer“ könnten sich auch manche positiven Möglichkeiten in ihr Gegenteil verkehren.

Es wäre sicher lohnend, sinnvolle, positive und gut funktionierende Anwendungen des Einsatzes von Generativer KI in der Politik zu sammeln und auf einer Art „Demokratie-Plattform“ zugänglich zu machen; diese Plattform könnte auch im Zusammenhang mit KI-Literacy-Initiativen (siehe Abschnitt 6.2) eine Rolle spielen. Für den Erfolg dieser neuen Elemente des zukünftigen demokratischen Diskurses ist essentiell, dass sie nach europäischen/österreichischen Standards implementiert werden, also etwa mit verlässlichen, DSGVO- und Urheberrechtskonformen Daten trainiert und entsprechend kuratiert werden. Weiters sollte die Gefahr eines neuen Digital Divide bedacht werden, also die vollständige Umstellung auf digitale Dienste, ohne dass alle Betroffenen diese auch schon nutzen, was zum Ausschluss mancher Bevölkerungskreise führen würde. Politischer Diskurs sollte wohl trotz aller Digitalisierung weiterhin unter Menschen stattfinden und nicht primär über KI-gesteuerte Bots.

[65] Allerdings entdeckte die Entwicklerin Lilith Wittmann erhebliche Sicherheitslücken in der App, wodurch Daten von über 100.000 Besuchen und 18.500 Wahlkampfhelfer:innen offengelegt wurden. Die App wurde vorübergehend offline genommen, und obwohl die CDU zunächst Anzeige gegen Wittmann erstattete, zog die Partei die Beschwerde später zurück. Der Vorfall führte zu Kritik seitens des Chaos Computer Clubs (CCC) und verdeutlichte das mangelhafte Sicherheitsmanagement der CDU (Hurz 2021).

4 RISIKEN GENERATIVER KI FÜR DIE POLITISCHE MEINUNGSBILDUNG

Kapitel 4 analysiert zunächst die spezifischen Risiken von Generativer KI im Zusammenhang mit dem politischen Diskurs allgemein (4.1) und von Deepfakes im Besonderen (4.2) sowie die Risiken von Microtargeting (4.3) und schließlich die Auswirkungen der zunehmenden Machtkonzentration im Bereich der KI-Entwicklung (4.4). Daran anschließend werden sogenannte hybride Bedrohungen (4.5) und solche durch Cyberkriminalität (4.6) fokussiert und ein Zwischenfazit gezogen (4.7).

4.1 DISKURSVERZERRUNG ALLGEMEIN

Wie einleitend dargelegt, ist eine wesentliche Voraussetzung für eine funktionierende Demokratie der möglichst offene, unbeeinflusste und transparente Diskurs über politische, d. h. zur Entscheidung anstehende Themen. Eine Verzerrung dieses Diskurses durch Des- oder Misinformation, egal ob absichtlich oder nicht, ist aus demokratiepolitischer Perspektive zu vermeiden. Generative KI hat jedoch das Potenzial, nicht nur bereits bestehende Verzerrungen zu verschärfen (4.1.1), sondern auch spezifisch neue hinzufügen. In diesem Abschnitt beschäftigen wir uns allgemein mit diesem Potenzial (4.1.2), im folgenden Abschnitt (4.2) dann mit den besonderen Risiken durch Deepfakes.

4.1.1 VERSCHÄRFUNG BESTEHENDER VERZERRUNGEN

Generative KI könnte bereits heute bestehende Probleme der Diskursverzerrung in den Sozialen Medien verschärfen (Karmasin et al. 2024). Als komplexes soziotechnisches System ist die Generative KI in die Transformation der Kommunikation und der Medienlandschaft durch die Digitalisierung eingebettet. Zum Verständnis über die Folgen von Generativer KI im politischen Kontext müssen nicht nur die Funktionsweise, die soziale Praxis und die Anwendungen dieser KI-Technologie, sondern auch die Wechselwirkungen mit dem Strukturwandel durch die digitale Kommunikation betrachtet werden. Dabei ist insbesondere auf neue oder verstärkende Effekte zu fokussieren.

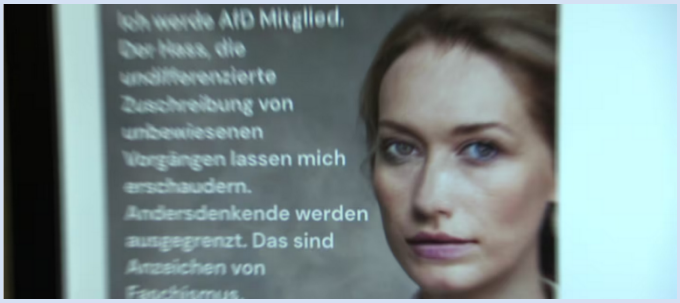
Demokratisierung der Kommunikation

Ein besonderer Aspekt digitaler Kommunikation ist die Möglichkeit, Informationen niedrigschwellig zu erstellen und zu veröffentlichen, die danach einfach und massenhaft weiterverbreitet werden können. Durch das Weitersenden, Liken, Teilen und Verlinken durch Nutzer:innen von Plattformen können so in kurzer Zeit große Reichweiten erlangt werden (wobei die Algorithmen, die die Nachrichten sortieren, eine wichtige Rolle spielen; deren Ziel ist es, im Falle der großen kommerziellen Plattformen, die Interaktionen mit den einzelnen Beiträgen

zu erhöhen, um die Verweildauer der Nutzer:innen auf den Plattformen zu verlängern[66]). Die digitale Kommunikation führt so zu einer *Demokratisierung der Kommunikation*, da Informationen und Äußerungen von allen erstellt und auf Plattformen in sozialen Medien verbreitet oder kommentiert werden können. Dadurch ergeben sich im politischen Kontext beispielsweise neue Möglichkeiten der Interaktion von Bürger:innen und gewählten Repräsentant:innen. Politische Akteure nutzen deshalb soziale Netzwerke zur politischen Kommunikation und Selbstdarstellung (Jungherr 2023) (siehe dazu auch Kapitel 3). Allerdings haben nicht alle dasselbe Gewicht, da die sozialen Medien aufgrund ihrer Algorithmen nicht alle gleichbehandeln.

Aufmerksamkeits-ökonomie

Andererseits verlieren klassische Medienformate wie Zeitung, Fernsehen oder Radio mit ihren journalistischen Gatekeepern zunehmend an Bedeutung. Die Meinungsbildung findet immer mehr durch Inhalte und Informationen in Sozialen Medien statt, bei denen die Plattformbetreiber für die Moderation verantwortlich sind. Weiterhin entsteht durch diese digitalen Möglichkeiten der Kommunikation eine wachsende Flut an Informationen, der man täglich ausgesetzt ist (sog. „Infodemie"). Unterschiedliche Akteure stehen im Wettbewerb um die *begrenzte Aufmerksamkeit* der Nutzer:innen und arbeiten verstärkt mit visuellen und auffallenden Inhalten, die besonders schnell rezipiert werden (Karmasin et al. 2024, S. 16f). Das bewirkt eine stärker emotionale und affektive Kommunikation. Gerade vor Wahlen ist eine verschärfte Konkurrenz um die Aufmerksamkeit von Wähler:innen zu beobachten.



Quelle: *swr.de/swraktuell/baden-wuerttemberg/stuttgart/afd-deepfakes-wahlkampf-100.html*

Für Wahlkampfzwecke verwendete die Partei AfD Göppingen in Deutschland ein synthetisch generiertes Gesicht einer fiktiven Person, konkret einer blonden Frau mit Doktortitel, die in einem Post auf Instagram Dr. Stefanie Müller genannt wird und erklärt, warum sie Mitglied der AfD geworden ist.

Box 2: AfD-Wahlplakat mit KI-generierter Bürgerin

[66] Kontroversielle Inhalte werden durch die Algorithmen häufig weiter nach oben gereiht als konsensuale. Das begünstigt besonders die Verbreitung von Desinformation, siehe unten.

Erstellung von irreführenden, schädigenden und polarisierenden Inhalten

Durch die Digitalisierung der Kommunikation können nutzergenerierte Inhalte auf Plattformen individuell erstellt und verteilt werden. Diese Inhalte können nicht nur unbeabsichtigt falsche und ungenaue Informationen (Misinformation) oder absichtlich irreführende und Informationen enthalten (Desinformation), sondern auch potenziell strafrechtlich relevante Inhalte darstellen.[67]

Desinformationskampagnen

Viele Beobachter:innen beschäftigen sich mit Desinformationskampagnen. Villar García et al. kommen etwa zum Schluss, dass „[i]n the realm of communication, disinformation campaigns are the most frequent component" (2021, S. 132). Dumbrava beschreibt das dadurch entstehende Problem wie folgt:

> *„Obwohl es immer mehr Belege dafür gibt, dass Menschen in erheblichem Umfang mit politischer Desinformation im Internet konfrontiert sind, ist es schwierig, die tatsächlichen Auswirkungen von Desinformation auf ihre Ansichten und Präferenzen zu beurteilen. Wenngleich die Reichweite und die Auswirkungen von Desinformation offenbar überschätzt wurden, gibt es jedoch Hinweise auf negative Auswirkungen in bestimmten Kontexten und auf bestimmte Gruppen. Durch Desinformation können Wählerinnen und Wähler gewonnen, aber auch verwirrt werden, und Bürgerinnen und Bürger können dazu bewegt oder davon abgehalten werden, sich an Wahlen zu beteiligen. Dies kann unter bestimmten Umständen den Ausgang einer Wahl beeinflussen (Verfälschung von Wahlergebnissen)."* (Dumbrava 2021, S. I f)

Hassrede

Zu den illegalen Inhalten zählen insb. Hassreden (hate speech) oder pornografische Darstellungen ohne die Einwilligung der Betroffenen (siehe Abschnitt 4.2). Unter Hassrede versteht man Inhalte oder Nachrichten, die sich gegen ein Individuum oder eine Gruppe richten, indem Identitätsmerkmale wie beispielsweise Geschlecht, Ethnie, Hautfarbe, Religion oder Nationalität des Individuums bzw. der Gruppe als negativ und unerwünscht dargestellt und die Abwertung eines Individuums bzw. einer Gruppe beabsichtigt wird. Dadurch stellt Hassrede nicht nur einen konfrontativen Debattenstil dar, sondern kann auch zu körperlicher Gewalt führen (Rudnicki/Steiger 2020; Dreyer et al. 2021). Der Einsatz von Hassrede und Desinformation kann Politiker:innen direkt schädigen und indirekt den demokratischen Diskurs und die Meinungsbildung beeinflussen.

Manipulierte Verbreitung von digitalen Inhalten: Fake Accounts

Zur Verbreitung von Inhalten in Sozialen Medien können Netzwerke von Accounts manuell betrieben werden, aber auch automatisiert arbeiten und eine Priorisierung des Inhalts bewirken. Oft werden diese Accounts unter falschen Angaben angelegt („Fake Accounts"). Das automatisierte Teilen und Verlinken stellt einen „manipulativen Missbrauch des Wechselseitigkeitsprinzips (reciprocity abuse) dar" (Karmasin et al. 2024, S. 26):

> *„In Hinblick auf demokratische Grundprinzipien wie Pluralismus, Meinungsvielfalt und freie Meinungsbildung sowie den Umstand, dass politische Relevanz nicht allein von Quantitäten abhängt, sind diese Mechanismen kritisch zu beurteilen."*
> (ebenda)

Noch weitergehend ist das sog. Microtargeting, also die gezielte Ansprache besonderer Bevölkerungs- oder Interessengruppen (siehe Abschnitt 4.3). Da die Steuerung der Verbreitung von Inhalten und Moderation von Plattformen i.d.R. mittels automatisierter Selektions- und Empfehlungs-Algorithmen erfolgt und nicht der öffentlichen Kontrolle oder dem nationalstaatlichen Rechtszugriff unterliegen, erhalten sie als Informationsintermediäre eine machtvolle Rolle in po-

[67] Das Phänomen der Fake News ist schon länger beobachtet worden; eine empirische Analyse kam zum Schluss, dass dessen Verbreitung 2018 noch eher marginal war (Fletcher et al. 2018).

litischen Diskursen und der Meinungsbildung (Allea 2021), siehe dazu auch Abschnitt 4.4. Es besteht weiterhin die Gefahr einer gesellschaftlichen Fragmentierung und Polarisierung (Dumbrava 2021).

Fragmentierung der Kommunikation (Echokammern, Filterblasen)

Soziale Medien und Plattformen bieten im digitalen Raum verschiedene Möglichkeiten, dass sich Menschen mit ähnlichen Interessen treffen und austauschen können. Beispielsweise gibt es geschlossene Foren, in denen Moderator:innen entscheiden, welche Informationen erlaubt sind, welche Fragen blockiert werden oder welche Personen aus dem Forum ausgeschlossen werden. Damit werden die Möglichkeiten des Widerspruchs oder des Austauschs mit anderen reduziert und die wiederholte Exposition ähnlicher Informationen kann zu einem Bestätigungseffekt führen. Es entsteht eine Spirale *aus sog. Echokammern, Filterblasen* und selektiver Wahrnehmung. Damit ist gemeint, dass es durch die Algorithmen der Plattformen eine Tendenz zur Abschottung von kleineren Gruppen mit ähnlichen Meinungen kommt, die zu einer Verengung der Weltsicht mit einem Potenzial zu Bestätigungsfehlern führen kann (Echokammern[68]) und die durch Suchmaschinen zur Verfügung gestellte Information auf Basis von bisherigen Suchanfragen und individuellen Profilen soweit gefiltert wird, dass nicht alle Informationen für alle zugänglich sind (Filterblasen, siehe Pariser 2012). Weiters kann es in den neuen Kommunikationsräumen zu unterschiedlichen Wahrnehmungen in Hinblick auf den Wahrheitsgehalt von Beiträgen und einen sehr konfrontativen Debattenstil kommen. Darüber hinaus ist das Phänomen der gezielten Desinformation und bewusster Meinungsmanipulation durch sog. Trollfabriken und andere Akteure beobachtet worden.

Die verstärkende Rolle der Generativen KI

Die beschriebenen Phänomene könnten durch Generative KI in mehrfacher Hinsicht verstärkt werden:

- Insbesondere wird nun zusätzlich die synthetische und personalisierte Form der Erzeugung und Vermittlung von Information ermöglicht, da die Programme zur Erstellung von Texten, Bildern oder Videos einer breiten Öffentlichkeit zugänglich sind. Mögliche Manipulationen und gezielte Irreführungen werden so deutlich einfacher und ressourcengünstig verfügbar.[69]
- Traditionelle journalistische Strukturen werden durch die ubiquitäre Verfügbarkeit dieser Tools weiter verdrängt. Damit wird es zunehmend schwieriger, Verantwortlichkeiten für Informationen und Nachrichten zu bestimmen und Qualitätsprüfungen fallen weg.
- Es kommt zu einem Wettlauf zwischen Anwendungen, die in der Lage sind, künstlich generierte oder veränderte Inhalte zu erkennen, und den Möglichkeiten einer perfekten Erzeugung (siehe Abschnitt 2.3). Kritischer Journalismus, der Fakten und Quellen prüft und sich auf ein Berufsethos stützt, gerät noch stärker unter Druck.

---

[68] Es sei an dieser Stelle angemerkt, dass die Konzepte der Echokammern und Filterblasen empirisch umstritten sind (vgl. bspw. Dubois/Blank 2018). Siehe dazu auch die Ausarbeitung des Wissenschaftlichen Dienstes des Deutschen Bundestags (2022), ***bundestag.de/resource/blob/898208/396d70db93fbc68bca40726b4d5308db/WD-10-007-22-pdf-data.pdf***.

[69] Insbesondere die Erstellung manipulierter Bilder oder täuschend echter Stimmen durch Deepfakes (dazu im nächsten Abschnitt 4.2 genauer) können zu problematischen Auswirkungen auf die Prozesse der politischen Meinungs- und Willensbildung führen (Thiel/Kailitz 2024). Das liegt daran, dass insbesondere Bilder oder Videos stärker emotionale Reaktionen auslösen und unmittelbarer wirken als Texte. Außerdem werden derartige Inhalte oftmals als glaubwürdiger eingeschätzt als Texte.

4.1.2 NEUE RISIKEN DURCH GENERATIVE KI

Es sind aber auch zusätzliche spezifische Folgen der Generativen KI und die besondere Bedeutung der Risiken durch Mis- und Desinformation hervorzuheben.

Systematische Verzerrungen durch die verwendeten Daten im Trainingsprozess des Maschinellen Lernens (Bias)

Die Funktionsweise Generativer KI-Technologien beruht auf der Erzeugung digitaler Inhalte durch die Verarbeitung großer Datenmengen in einem Trainingsprozess und auf einer Vielzahl aktiver und kontinuierlich zu fällender Moderationsentscheidungen. Für die Auswahl an Daten, den Trainingsprozess und das Feintuning sind die Produzenten der Technologie verantwortlich, wodurch diese eine dominante Gatekeeper-Rolle erhalten (siehe Abschnitt 4.4). Für die Generierung neuer Inhalte werden bei allen Verfahren zunächst große, unbereinigte Datensätze aus dem öffentlichen Internet automatisiert gesammelt. Diese Daten können auch ethisch problematische, unausgewogene oder unkorrekte Informationen enthalten. Das kann beispielsweise der Fall sein, wenn soziale Gruppen in dem Trainingsmaterial unterschiedlich repräsentiert sind. In einem weiteren Schritt werden dann bestimmte Muster erkannt und wiederholt, wodurch dann auch verzerrte Wiedergaben entstehen können.

Mittelfristig: Einschränkungen der Diversität

Jüngst wurde beschrieben, dass es zunehmend auch zu unerwarteten Problemen bei der Fortentwicklung von Generativer KI kommen kann, wenn diese nämlich in Zukunft überwiegend mit KI-generierten Daten trainiert werden. Martínez et al. (2024) benennen das Phänomen als „Krankheit" (Model Autophagy Disorder oder MAD): Das Verarbeiten von Daten aus einem geschlossenen Kreislauf (also KI-generierten Texten oder Bildern) und den darin enthaltenen Fehlern kann im Lernprozess einer KI zum Zerfall des KI-Modells führen. Konkret würden die Ergebnisse aufgrund der zugrundeliegenden Wahrscheinlichkeitsalgorithmen immer ähnlicher, immer weniger kreativ, was zu einer Einschränkung der Diversität und Meinungsvielfalt führen könnte.[70]

Verzerrende Effekte durch Trainingsdaten

Mehrere Autor:innen konnten beispielweise sexistische oder rassistische Äußerungen oder Darstellungen durch Textgeneratoren oder Bildgeneratoren nachweisen (Dale 2021; Davey 2022; Stokel-Walker/Van Noorden 2023; Ananya 2024). Damit werden explizite Werte wie Diversität und Inklusivität unterlaufen, die gemäß einer Empfehlung der UNESCO zur Ethik von KI respektiert werden sollten (UNESCO 2022). Der verzerrende Effekt kann zusätzlich verstärkt werden, wenn die erzeugten Inhalte selbst wieder Grundlage künftiger Trainingsdaten werden. Somit besteht die begründete Befürchtung, dass Verzerrungen und Stereotype durch Generative KI nicht nur dauerhaft aufrechterhalten, sondern künftig noch verstärkt werden, wenn es keine effektive Gegensteuerung gibt.[71] Dies kann zu Diskriminierungen oder auch zu Irreführungen oder unausgewogenen Inhalten führen. Bei ChatGPT wurde beispielsweise ein tendenziell links-liberaler politischer Einschlag in den erzeugten Texten identifiziert (Rozado 2023). Eine stärkere Nutzung dieses Systems beispielsweise in der wissenschaftlichen Politikberatung könnte dem Anspruch auf Überparteilichkeit entgegenstehen (Albrecht 2024, S. 25).

---

[70] *faz.net/pro/d-economy/kuenstliche-intelligenz/wenn-kuenstliche-intelligenz-krank-wird-19902462.html*.

[71] Wie schwer das ist, bzw. andere unerwünschte Nebeneffekte erzeugt, zeigt eindrücklich, als Google jüngst People of Color in Nazi-Uniformen steckte: *nytimes.com/ 2024/02/22/technology/google-gemini-german-uniforms.html*.

Beispiel für Bias

Dies verdeutlicht insbesondere, dass Technologien generell nicht wertfrei sind, sondern durch Designentscheidungen der Entwickler:innen geprägt werden. Im Detail wurde dies für die Erstellung von Datenbanken für den Bildgenerator Stable Diffusion beschrieben. Ein Team des Stanford Internet Observatory hatte darauf hingewiesen, dass die zugrundeliegende Datenbank LAION-B auch Abbildungen von sexueller Gewalt an Kindern beinhalten könnte (Beuth et al. 2024). Die Anbieter stellten daraufhin einen bereinigten Datensatz zur Verfügung. Die meisten Anbieter von KI-Modellen, wie z. B. openAI, geben jedoch nur noch wenig über die von ihnen verwendeten Trainingsdaten preis.

Verbreitung von Misinformation durch die mangelnde Genauigkeit von KI-Textgeneratoren in Bezug auf Fakten (Reproduzierbarkeit)

Generative KI kann auch ohne Täuschungsabsicht der Entwickler:innen der Technologie oder Ersteller:innen von Inhalten falsche, ungenaue und zugleich überzeugende Information erzeugen. Aufgrund der hohen Reichweite kann der potentielle Schaden somit erheblich sein.

Die Ausgabe von Textgeneratoren wie ChatGPT wird nämlich mit einer Wortfolgestatistik erstellt, indem die Wahrscheinlichkeit eines aufeinanderfolgenden Wortes (oder eines Wortteils, Token genannt) in dem zum Training verwendeten Textkorpus berechnet wird (Wolfram 2023) (siehe schon Abschnitt 2.1). Diese Art von linguistisch plausiblen Texten basiert nicht auf Semantik, Wissen oder ethischer Reflexion (Taecharungroj 2023). Außerdem beruhen sie nicht in erster Linie auf Fakten und können sogar frei erfundene Inhalte oder Quellen enthalten (sogenannte „Halluzination“) (Ji et al. 2023).

„Halluzinationen“

> *„Generative AI tools may produce „hallucinations“— erroneous responses that seem credible. One reason hallucinations occur is when a user requests information not in the training data. Additionally, a user could use AI to purposefully and quickly create inaccurate or misleading text, thus enabling the spread of disinformation.“* (STAA 2023)

Weiterhin dürfen keine Erwartungen an eine Abstraktion oder eine logische Problemlösung gestellt werden (Bender/Koller 2020). Es handelt sich zwar um leistungsfähige Sprachmodelle, aber eben nicht um Wissensmodelle (Heil 2023). Es ist daher wichtig, ausreichend Kenntnisse über die Funktionsweise dieser Technologien zu vermitteln und die Nutzung von KI-Textgeneratoren transparent zu kennzeichnen, um möglichen Misinformationen vorzubeugen.

Desinformation, Propaganda durch generierte synthetische Inhalte

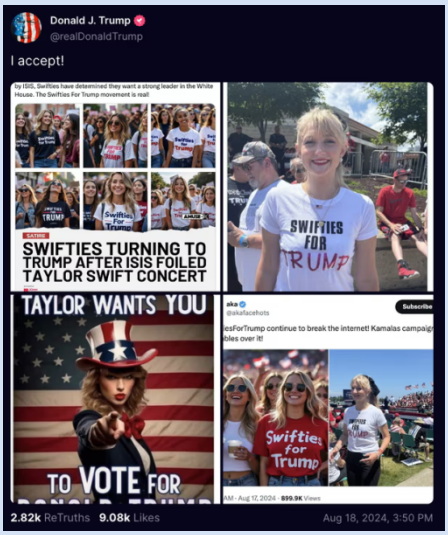
Generative KI kann auch für eine gezielte Desinformation eingesetzt werden, also der Produktion von Falschinformationen mit der Absicht zu manipulieren und zu täuschen. Der zunehmende Einsatz von Textgeneratoren in den traditionellen wie auch Sozialen Medien ist mit der Zunahme von gezielt oder unbewusst verbreiteter Desinformation verbunden (Ananthaswamy 2023; Atleson 2023). Damit wachsen die Chancen Einzelner, aber auch strategisch agierender Gruppen, die demokratische Öffentlichkeit zu täuschen und das Vertrauen in demokratische oder journalistische Akteure zu beschädigen.

Deepfakes im Superwahljahr 2024

Im sog. Superwahljahr 2024 hat das US-Magazin WIRED die Nutzung von Generativer KI im Kontext demokratischer Wahlen weltweit gesammelt und 78 Beispiele finden können.[72] In der Folge haben Forscher:innen der Princeton-Universität diese genauer betrachtet, unter besonderem Augenmerk auf drei Aspekte:[73] Gab es einen Vorsatz der Irreführung, handelte es sich um Deepfakes oder Cheapfakes und letztlich, welche Kosten würden anfallen, wollte man diese mit anderen Methoden als Generativer KI erstellen (Photoshop, Schauspieler:innen

[72] *wired.com/story/generative-ai-global-elections/*.
[73] *aisnakeoil.com/p/we-looked-at-78-election-deepfakes*.

oder auch geschicktes Schneiden eines Videos). Das Ergebnis: Oft konnten sie keinen irreführenden Charakter feststellen (etwa indem ein Kandidat in einer Sprache, die er nicht beherrscht, sprach[74]). Viele andere Beispiele waren außerdem gar keine Deepfakes, sondern eben Cheapfakes. Und diese wären vermutlich auch nicht signifikant teurer als KI-generierte Fakes. Zusammen schlussfolgern die Forscher:innen, dass die Schwemme an Deepfakes bislang ausgeblieben ist. Dennoch bleiben einige Frage offen, ob dies einerseits nicht auch ein Resultat der erhöhten Sensibilität für deren Risiko in demokratischen Prozessen war und ob nicht auch Generative KI eine gewisse Lernkurve hat und sich die Situation in naher Zukunft noch deutlich ändern könnte. Letztlich bleibt auch die methodische Frage ungewiss, ob nicht u. U. in nicht-öffentlichen Kanälen Deepfakes verbreitet wurden, die WIRED einfach nicht bekannt waren oder besonders gut gemachte Deepfakes einfach nicht als solche erkannt wurden.



Quelle: *theguardian.com/technology/article/2024/aug/24/trump-taylor-swift-deepfakes-ai*

Donald Trump postete im US-Präsidentschaftswahlkampf 2024 KI-generierte Fotos von Popstar Taylor Swift und von ihren Fans, die angeblich Donald Trump unterstützen. Taylor Swift hatte jedoch mehrfach erkennen lassen, dass sie eine Unterstützerin der demokratischen Kandidatin ist.

Box 3: Taylor-Swift-Fans pro Trump

Ein anderer Bericht widerspricht der Einschätzung der Princeton-Forscher:innen und gibt durchaus Anlass zur Sorge. Die US-Organisation Newsguard dokumentierte im Frühjahr 2023 fast fünfzig Websites in sieben unterschiedlichen Spra-

[74] Diese Bewertung könnte freilich auch in Zweifel gezogen werden, denn es ist durchaus irreführend, Fähigkeiten vorzugeben, die man nicht besitzt. So wäre es etwa in einem Land wie Indien, in dem viele Sprachen gesprochen werden, für einen landesweit antretenden Kandidaten durchaus eine hervorragende Eigenschaft, wenn er viele Landsleute in der eigenen Sprache ansprechen könnte.

chen, „die augenscheinlich alle in großen Mengen KI-generierte Texte als vermeintliche Nachrichtenartikel präsentierten“ (Stöcker 2024, S. 412). In einem Artikel wurde sogar über den angeblichen Tod von US-Präsident Joseph „Joe“ Biden berichtet (Holland 2023). Neben der inhaltlichen Irreführung durch eine falsche Nachricht liegt hier eine Irreführung durch eine fehlende Kennzeichnung der generierten Inhalte vor.

Im politischen Kontext handelt es sich insbesondere um Desinformation zur Beeinflussung von Wahlen, aber auch um die Nutzung von Generativer KI für Kampagnen durch Politiker:innen und Parteien (Chowdhury 2024, S. 413; Stöcker 2024). Letztere bewirken zwar meist keine konkreten Rechtsverletzungen, sind aber trotzdem ethisch problematisch. Die Risiken durch Deepfakes für den politischen Diskurs werden im folgenden Abschnitt vertieft betrachtet.

Verschleierung der Akteure und Quellen

Durch den Einsatz von Generativer KI kann auch systematisch verschleiert werden, wer eigentlich agiert, also eine bestimmte Nachricht verbreitet oder aus welcher Quelle die ursprünglichen Inhalte stammen. So könnte etwa ein noch nicht veröffentlichter Gesetzesentwurf mittels Generativer KI auf einfache Weise so verändert werden, dass er eine andere regulative Aussage bekommt, und anschließend würde dieser sozusagen „geleakt“, sprich mittels Bots weit gestreut. In diesem Fall scheint es dann so, als ob der Entwurf von einem offiziellen Akteur stammt (z. B. der Bundesregierung oder der EU-Kommission).

Chatbot als politischer Berater

Bei Anwendung fortgeschrittener Generativer KI in Form von personalisierten Chatbots könnte zwischen User:in und KI eine Art Vertrauensverhältnis entstehen, sodass nicht nur – wie bereits empirisch bekannt – intime private Angelegenheiten mit dem scheinbaren „Freund“ besprochen werden, sondern zunehmend auch politische Fragen bis zur Beratung bei Wahlentscheidungen. Damit wird offensichtlich, welche potenzielle politische Macht Generative KI haben kann und wie wichtig es daher ist, dass diese nur auf objektive, geprüfte, nicht desinformierende Quellen ohne Bias zurückgreift und von niemandem im Hintergrund sozusagen „gelenkt“ wird.

4.2 DISKURSVERZERRUNG DURCH DEEPFAKES

Disruptives Potenzial von Deepfakes

Im vorigen Abschnitt (4.1) wurden die Risiken von Generativer KI im politischen Bereich allgemein erörtert. Ein spezifisches KI-generiertes Phänomen, nämlich Deepfakes, hat darüber hinaus disruptives Potenzial, da insbesondere Politiker:innen eine potenzielle Zielgruppe sind, vor allem dann, wenn sie nicht über die Ressourcen verfügen, um ihre Online-Präsenz zu schützen. Für die Gesellschaft besteht durch Deepfakes darüber hinaus die große Gefahr, dass das Vertrauen in das Informationsumfeld und den politischen Prozess untergraben wird. Dieser Abschnitt konzentriert sich daher auf die spezifischen Risiken von Deepfakes.

Drei Ebenen betroffen: Person, Organisation, Gesellschaft

Die Folgen der Anwendungen der Deepfake-Technologie ohne ausreichende Kennzeichnung oder ohne Einverständnis der dargestellten Personen sind multidimensional und ergeben sich einerseits direkt für die täuschend echt dargestellte Person (Mikro-Ebene), weiterhin auch für die zugehörige Organisation oder Institution beispielsweise Parteien oder Medien (Meso-Ebene) und schließlich auch auf einer systemischen gesellschaftlichen Makro-Ebene hinsichtlich des Einflusses auf politische Diskurse. Oft zeigen sich die Folgen nicht nur auf einzelnen Ebenen, sondern aufgrund kaskadierender Effekte in allen drei Betroffe-

nen-Ebenen. Beispielsweise kann ein Deepfake, der die Diskreditierung eines/r Politiker:in zum Ziel hat, zugleich die zugehörige Partei schädigen und auf der gesellschaftlichen Ebene die Beeinflussung einer Wahl beabsichtigen (van Huijstee et al. 2021).

Fünf Angriffstypen

Einen Überblick über unterschiedliche Arten von Deepfakes anhand des Angriffstyps und anhand der drei Betroffenenebenen geben Karaboga et al. (2024, S. 277ff). In dieser TA-Swiss-Studie werden fünf grundlegende Typen unterschieden:

- Gefälschte Aussagen oder Handlungen
- Social Engineering, d. h. das Ausnutzen menschlicher Eigenschaften wie Hilfsbereitschaft, Vertrauen oder Angst, um Personen zu manipulieren
- Überwindung von Sicherheitsmaßnahmen (Authentifizierungssysteme)
- Synthetische Social Botnets, d. h. Computerprogramme, die menschliche Identitäten vortäuschen und im Internet automatisiert kommunizieren
- Deepfakes von Objekten und Situationen (vollständig synthetische Inhalte)

Elf Szenarien

Weiters identifizieren Karaboga et al. (2024, S. 281) anhand der bisher beobachteten Beispiele insgesamt elf unterschiedliche Szenarien, wie Deepfakes im politischen Kontext Verwendung finden können (siehe Tabelle dort S. 283f), von „Einschüchterung eines Politikers“ über „Rufschädigung einer Partei“ bis zu „Beeinflussung einer Wahl“, systematisiert nach der Betroffenenebene, dem Angriffstyp und Kreis der Adressat:innen. In diesem Abschnitt fokussieren wir auf Deepfake-Anwendungen mit potentiellen Auswirkungen auf den politischen Diskurs und das politische System der Demokratie. Spezifische strafrechtsrelevante Cyberangriffe wie Erpressung, Identitätsdiebstahl oder die Überwindung von Sicherheitsmaßnahmen werden im Detail in Abschnitt 4.6 behandelt. Wir untersuchen in der Folge diese drei unterschiedlichen Betroffenen-Ebenen und ordnen Beispiele entsprechend zu.

4.2.1 INDIVIDUELLE EBENE (MIKRO-EBENE)

Diskreditierung von Personen führt zu psychologischen und finanziellen Folgen

Auf der Mikro-Ebene werden gezielt und direkt einzelne Personen diskreditiert oder Zielpersonen vorsätzlich getäuscht und manipuliert. Sind die visuellen Inhalte ohne Zustimmung der Personen generiert, liegt in der Regel eine Verletzung von Persönlichkeitsrechten und somit eine illegale Handlung vor. Werden Personen bei Handlungen oder mit Äußerungen gezeigt, die diese in Wirklichkeit nie gemacht haben, so kann es zu Rufschädigungen führen. Die Täter:innen bleiben meist anonym und versuchen die Opfer zu bedrohen, einzuschüchtern oder zu erpressen (vgl. Abschnitt 4.6). Somit ergeben sich sowohl psychologische als auch finanzielle Schäden für die Betroffenen, die in der Regel strafrechtlich relevant sind (van Huijstee et al. 2021). Während anfangs hauptsächlich Schauspieler:innen und bekannte Personen dargestellt wurden, zu denen ausreichend audiovisuelles Material im Internet vorliegt, reichen heute auch einzelne Fotos von Personen zur Erzeugung von Deepfakes aus.

Kaskadierende Effekte bei Deepfakes mit Politiker:innen

Als problematisch sind neben dem klassischen Gesichtstausch in Videosequenzen oder Fotos insbesondere Gesichtsmanipulationen, d. h. die Übertragung der Mimik eines Gesichts auf ein anderes Gesicht. Mit dieser Technologie können beispielsweise Politiker:innen Aussagen in den Mund gelegt werden, die sie nie getätigt haben. Aber auch Journalist:innen können durch Deepfakes diskreditiert werden, indem manipulierte Interviews oder falsche Interviewpartner:innen dargestellt werden (Karaboga et al. 2024, S. 211). Das kann nicht nur Folgen

für die Person auf der Mikro-Ebene haben, sondern aufgrund kaskadierender Effekte auch die repräsentierte Partei oder die Medien diskreditieren (Meso-Ebene) und damit indirekt auch einen Einfluss auf die Meinungsbildung in der Gesellschaft auf der Makro-Ebene darstellen.

Beispiele von diskreditierenden Deepfakes von Politiker:innen

Es gibt bereits etliche Fälle von Deepfakes von Politiker:innen: So wehrte sich etwa Italiens Ministerpräsidentin Giorgia Meloni vor Gericht gegen gefälschte Sexvideos, bei denen ihr Gesicht auf den Körper von Pornodarstellerinnen geschnitten wurde. Die Filme waren nach Angaben der Ermittler 2020 monatelang im Internet abrufbar und wurden auch millionenfach angesehen (Spiegel Ausland 2024). Ein weiteres Beispiel für die Diskreditierung einer Politikerin ist das mit einfachen Mitteln erzeugte Video der früheren Sprecherin des Repräsentantenhauses der USA, Nancy Pelosi. Mit einer langsameren Abspielgeschwindigkeit wird hier Nancy Pelosi als vermeintlich betrunken gezeigt. Das Video wurde nicht mittels KI erzeugt und wird deshalb genau genommen als sogenanntes „Cheapfake" („cheap" = engl. für „billig") und nicht als Deepfake eingeordnet. Das manipulierte Video wurde durch republikanische Politiker:innen zur Wahlmanipulation verbreitet (Harwell 2019). Ein aktuelles Beispiel aus dem US-Wahlkampf ist das auch von X- (ex-Twitter-) Chef Elon Musk (neben vielen anderen Falschmeldungen weiterverbreitete Video über Präsidentschaftskandidatin Kamela Harris, in dem sie angeblich behauptet, dass sie nur wegen ihres Geschlechts und ihrer Hautfarbe Kandidatin wäre und daher alle Angriffe gegen sie sexistisch und rassistisch wären.[75] Ebenfalls durch eine KI-Attacke getroffen wurde Rumin Farhana. Ende 2023 war die führende Oppositionelle aus Bangladesch auf unterschiedlichen Plattformen so dargestellt, dass in dem mehrheitlich muslimischen Land eine Empörung ausgelöst wurde, obwohl klar wurde, dass es sich um einen Deepfake handelt.[76]

4.2.2 ORGANISATIONS-EBENE (MESO-EBENE)

Erfolgt der Angriff nicht explizit auf eine Privatperson, sondern beispielsweise auf eine entscheidungsbefugte oder repräsentative Person einer Organisation oder auf diese selbst, so wirkt sich der Schaden auch auf die Organisation und nicht nur auf die betroffene Person selbst aus. Bekannte Beispiele hierfür sind Audio-Deepfakes mit täuschend echten Stimmen von Vorgesetzten, die um eine Überweisung oder die Ausgabe von Passwörtern bitten (Voice spoofing, Social Engineering im Rahmen eines Betrugs, siehe auch Abschnitt 4.6).

Diskreditierung von Journalismus und Medien führt zu Glaubwürdigkeitsverlust und schwächt die Demokratie

Dem Journalismus wird eine wichtige Rolle in einer Demokratie zugeschrieben, denn er stützt sich auf professionelle Normen zur kritischen Prüfung von Informationen. Durch die journalistische Arbeit in der Berichterstattung der Medien sollen den Bürger:innen vertrauenswürdige Informationen über relevante Themen für den politischen Diskurs und zur Meinungsbildung zur Verfügung gestellt werden (Burkart 2021). Medien erfüllen damit wichtige Funktionen in der politischen Kommunikation, wie beispielsweise das Agenda Setting oder Framing (Schünemann 2022, S. 40). Deepfakes, die direkt die Medien diskreditieren, führen letztlich nicht nur zu einem Glaubwürdigkeitsverlust des Journalismus und

[75] *counterhate.com/research/musk-misleading-election-claims-viewed-1-2bn-times-on-x-with-no-fact-checks/*; siehe auch: *orf.at/stories/3366109/*.

[76] *deutschlandfunk.de/ki-wahlen-manipulation-kuenstliche-intelligenz-fake-news-deepfakes-100.html*.

der Institutionen der Medien, sondern auch zur Schwächung der Demokratie (Godulla et al. 2021).[77] Die aktuelle TA-Swiss-Studie hält dazu fest:

> *„Sowohl die versehentliche Verbreitung von Deepfakes durch die Medien selbst wie auch die Diskreditierung von Medienorganisationen durch Dritte bergen das Risiko des Vertrauensverlustes gegenüber den Medien."* (Karaboga et al. 2024, S. 225)

Deepfakes können aber auch benutzt werden, um Journalist:innen anzugreifen, um kritische Berichterstattung zu verhindern (Posetti 2018; Karaboga et al. 2024, S. 210). Dies stellt eine Bedrohung der Medienfreiheit dar.

Beispiele von Deepfakes zur Diskreditierung von Journalismus

Eine gezielte Diskreditierung der Medien in Deutschland erfolgte etwa in einem Audio-Deepfake eines Sprechers der Tagesschau, der sich für angebliche Lügen in der Berichterstattung entschuldigt (Reveland/Siggelkow 2023). In einem anderen Deepfake wird ein bekannter ZDF-Moderator mit geklonter Stimme zu Werbezwecken missbraucht (Breithut 2023).

Diskreditierung von Parteien

Im politischen Kontext können synthetische, KI-generierte Fotos und Texte in gefälschten Social-Media-Konten und synthetischen Social Botnets verwendet werden, um gezielt Parteien zu diskreditieren und damit Wahlen zu manipulieren (van Huijstee et al. 2021).

Beispiele

Ein Beispiel im deutschsprachigen Kontext ist die Diffamierung der Partei der Grünen in der Schweiz im Kontext der eidgenössischen Wahlen 2023 durch ein Deepfake mit manipulierter Stimme der Nationalrätin Arslan (Fm1Today 2023). Hier ging es nicht in erster Linie um die Diffamierung der Person Arslan, sondern um die Partei – interessanterweise nicht versteckt, sondern offen, den die gefakte Nationalrätin sagte in dem Video:

> *„Ich bin Sibel Arslan von den Grünen und ich will, dass alle kriminellen Türken ausgeschafft werden. Bitte wählen Sie bei den eidgenössischen Wahlen SVP und schreiben Sie Andreas Glarner zweimal auf Ihre Liste. Dies sagt nicht die richtige Sibel Arslan, sondern eine KI-generierte Version der grünen Nationalrätin. Der Absender: der Aargauer SVP-Nationalrat Andreas Glarner."* (Fm1Today 2023)

Ein weiteres Beispiel ist das durch das Kunstkollektiv Zentrum für Politische Schönheit (ZPS) erstellte KI-generierte Video von Bundeskanzler Scholz zum Verbot der AfD. Scholz wird in dem Video in den Mund gelegt, dass die Bundesregierung anlässlich des fünftes Todestages Lübckes im Sommer 2024 ein Verbot der AfD beim Bundesverfassungsgericht beantragen werde (Monopol 2023). Das Video musste von den Plattformen gelöscht werden, da zudem das Markenrecht der Bundesregierung durch die Verwendung des sogenannten „Flaggenstabes" verletzt wurde. (Dieser besteht aus dem Bundesadler, der stilisierten deutschen Fahne und einem Schriftzug).

Deepfakes für Propaganda

Politische Akteure und Parteien nutzen aber auch selbst generative Methoden und Deepfake-Technologien zur politischen Kommunikation, Propaganda und Selbstdarstellung – wie auch das obige Beispiel aus der Schweiz zeigt (Jungherr 2023). Propaganda ist ein Mittel der strategischen Kommunikation, im traditionellen Verständnis insbesondere in totalitären oder autoritären Regimen und richtet sich meistens an die inländische Bevölkerung (Schünemann 2022, S. 34, siehe auch Abschnitt 1.2).

---

[77] Dies gilt übrigens auch für die Institution der bloggenden Einzeljournalist:innen oder Recherchekollektive mit hoher Reichweite, also auch für Vertreter:innen nicht-klassischer Medien. Auch wenn deren Seriosität kompromittiert wird, hat das Auswirkungen auf die Wahrnehmung und Glaubwürdigkeit der Medienlandschaft insgesamt.

4.2.3 GESELLSCHAFTLICHE UND SYSTEMISCHE EBENE (MAKRO-EBENE)

Deepfakes zur Erstellung von Desinformation und damit zur Diskursverzerrung

Deepfake-Technologien sind ein effektives und effizientes Instrument zur Erstellung von Desinformation, weil die generierten audiovisuellen Inhalte authentisch erscheinen und schneller als Textinhalte rezipiert werden. Das führt einerseits zu Täuschungen und Manipulationen, andererseits zu Verunsicherungen und Vertrauensverlust in Informationen, da man sich nicht mehr auf das verlassen kann, was man sieht oder hört. Die Erosion von Vertrauen in Information und Fakten wirkt sich somit indirekt auf die Meinungsbildung aus, die vertrauenswürdige Informationen als Basis für Entscheidungen benötigen. Durch das Ineinanderwirken der automatisierten, aber auch manipulierten oder speziell zugeschnittenen Verbreitung von Informationen (siehe auch Abschnitt 4.3 zu Microtargeting) kann die Gefahr der Beeinflussung des Meinungs- und Willensbildungsprozesses in der Bevölkerung noch verstärkt werden (Dobber et al. 2020; Heesen et al. 2021). Deepfake-Technologien werden nicht nur für Propaganda und Wahlmanipulationen eingesetzt, sondern auch zum Schüren sozialer Unruhen, politischer Polarisierung oder Radikalisierung sowie zur Verunsicherung der Bürger:innen missbraucht. Dies stellt neben der inhaltlichen Diskursverzerrung eine ernstzunehmende Bedrohung für die Demokratie dar (Aïmeur et al., 2023; Ciancaglini, 2020; Europol, 2023).

Propaganda, Polarisierung, Radikalisierung, Verunsicherung

Für diese Art der Desinformation werden meist Deepfakes mit synthetischen Inhalten generiert, d. h. es werden fiktive Personen oder Situationen dargestellt. Ein Beispiel im aktuellen Krieg im Nahen Osten stellt das generierte Bild eines Mannes mit fünf Kindern dar. Das Bild wurde auf vielen Plattformen verteilt und erreichte hunderttausende Nutzer:innen in den sozialen Netzwerken. Expert:innen konnten zeigen, dass es sich um ein generiertes Foto handelt (Nicolaus 2023).

Stimmungsmache

Beispiele aus Wahlkämpfen

Für die eidgenössischen Wahlen in der Schweiz erstellte die FDP 2023 eine Plakatkampagne und verwendete dazu ein fiktives, generiertes Bild, das Klimaaktivist:innen auf einer Straße sitzend zeigt. Vor ihnen steht eine Ambulanz mit Blaulicht. Der Einsatz von KI wurde freilich auf allen Plakaten gekennzeichnet:

> *„Man wolle mit dem Plakat «Anpacken statt ankleben!» die «destruktiven Aktionen» von Klima-Aktivisten anprangern, so die FDP. Statt Strassenblockaden seien konstruktive Lösungen gefragt."* (Blick 2023)

Für weitere Beispiele aus Deutschland und Indien siehe Box 2 oben und die folgende Box 4.[78] Wenn die generierten Bilder, Videos oder Audios von den Kandidat:innen selbst erstellt werden, um diese als attraktiver und ansprechender erscheinen zu lassen, bezeichnet Stöcker (2024) diese Art von Irreführung auch als „Softfake". Daneben haben wir es auch mit Deepfakes mit Desinformations-Absichten zu tun, also um mittels Generativer KI veränderte visuelle Inhalte, die nicht von Politiker:innen selbst, sondern ohne deren Zustimmung von Täter:innen mit krimineller Absicht erstellt werden. Dies zeigt auch, dass die Risiken von Generativer KI ein kontinuierliches Spektrum zwischen unbeabsichtigter Täuschung, gezielter Desinformation und kriminellen Missbrauch darstellen.

[78] Für viele weitere Beispiele siehe diese Liste: *cs.princeton.edu/~sayashk/political-misinformation/WIRED-data.html*.

Quelle: *heise.de/news/KI-im-Wahlkampf-in-Indien-Wenn-der-tote-Vater-Wahlwerbung-fuer-den-Sohn-macht-9710648.html*

Im indischen Parlamentswahlkampf tauchten verschiedene Videos von Kandidat:innen auf, in denen diese etwa in verschiedenen Sprachen sprechen oder nostalgische Lieder singen und so präsentiert werden, wie es die Kandidat:innen selbst in der Realität nicht könnten. In einem Video macht angeblich ein verstorbener Vater für seinen Sohn Wahlkampf. In einem anderen (siehe Screenshot) diskutiert ein Kandidat mit einem Avatar seiner selbst.

Box 4: Indischer Politiker spricht im Video mit sich selbst

Deepfakes als politische Satire können irreführend sein

Deepfakes werden aber auch für politische Misinformation ohne direkte Schadensabsicht der Ersteller:innen verwendet. Im politischen Kontext ist hier insbesondere Satire zu berücksichtigen, die auf unterhaltsame Weise die Mächtigen kritisieren und damit Deliberation und den öffentlichen Diskurs anregen (Pawelec 2022). Es handelt sich zwar um von der Meinungsfreiheit geschützte Äußerungen, die aber durchaus ethisch problematisch sein können, da sie von irreführender Misinformation oftmals schwer zu unterscheiden sind. Ein Beispiel stellen die von vielen so genannten „Klitschko Deepfakes"[79] (die aber im technischen Sinne keine Deepfakes sondern schauspielerische Imitationen waren) dar, die im Sommer 2022 von den Darstellern Vovan and Lexus (Vladimir Kuznetsov und Alexey Stolyarov) erzeugt wurden, um Bürgermeister:innen aus unterschiedlichen Ländern, auch in Österreich, zu täuschen (Jonas/Marinov 2022; Maier/Schmid 2022).[80] Auch im aktuell anlaufenden Wahlkampf in Deutschland wurde bereits ein satirisch aufgemachtes Deepfake von Friedrich Merz geteilt.[81]

---

79 Siehe: *kleinezeitung.at/politik/aussenpolitik/ukraine/6162451/Russische-Komiker_KlitschkoFakeAnrufer-arbeiten-fuer-Plattform-von*; *kleinezeitung.at/politik/innenpolitik/6175481/Video-aufgetaucht_FakeKlitschko-an-Ludwig_Heben-Sie-die-Haende*; *deutschlandfunk.de/mediasres-fakes-in-der-politik-100.html*.

80 Siehe ein erstes Beispiel eines Satire-Deepfakes aus dem beginnenden österreichischen Nationalratswahlkampf, berichtet von Brodnig in *derstandard.at/story/3000000224778/ spricht-hier-kickl-oder-die-ki*.

81 *br.de/nachrichten/netzwelt/deepfake-von-merz-was-bedeutet-er-fuer-den-bundestagswahlkampf,UUCtkLv*.

Lügner:innen-Dividende: mögliche Leugnung echter Inhalte

Die (noch nicht bewältigte) Herausforderung, Medieninhalte zweifelsfrei zu authentifizieren, erlaubt es umgekehrt, jegliche kompromittierende Information zu leugnen. Das bedeutet, dass beispielsweise jede:r behaupten kann, dass ein Video gefälscht sei. Dieses Phänomen wird als Lügner:innen-Dividende („Liar's dividend") bezeichnet (Chesney/Citron 2018).

Beispiel: Festnahme George Floyd

Ein Beispiel für die Leugnung eines Beweisvideos stellt die Aufnahme des Polizeieinsatzes zur Festnahme von George Floyd in Minnesota dar, das von Verschwörungstheoretikern angezweifelt wurde mit dem Hinweis, dass Floyd schon früher gestorben sei und sein Gesicht nur als Deepfake verarbeitet worden sei, um Unruhen zu provozieren. Tatsächlich handelt es sich um einen Zusammenschnitt der Aufnahmen von Passant:innen und Überwachungskameras, in denen der Polizist sein Knie in den Nacken von George Floyd so lange gepresst hat, dass der Afroamerikaner daran kurze Zeit später verstarb. Die Tat hat eine Welle des Protests ausgelöst (Denkler 2021).

Epistemische Unsicherheit schwächt die Demokratie und nationale Sicherheit

Andererseits kann dieses Phänomen auch zur Folge haben, dass nach eingehender Unterrichtung der Bevölkerung über die Möglichkeiten und den Einsatz von Deepfakes auch die Skepsis gegenüber authentischen Quellen hervorgerufen wird (Chesney/Citron 2018), mit anderen Worten: dass also auch korrekte Informationen zunehmend in Zweifel gezogen werden könnten. Diese epistemische Unsicherheit schwächt in der Folge auch das Vertrauen der Bürger:innen in politische Institutionen, Behörden und besonders in die Medien. Dadurch wird nicht nur die Demokratie geschwächt, sondern können auch die nationale Sicherheit und die internationalen Beziehungen bedroht werden (siehe dazu Abschnitt 4.5).

4.2.4 ZWISCHENFAZIT

Desinformation, vor allem in Form von Deepfakes, ist eine Bedrohung insbesondere von liberalen Demokratien mit ihren Institutionen, da diese anfällig gegenüber Manipulationen von Informationen aufgrund der Medienfreiheit und Freiheit der Meinungsäußerungen sind. Dabei handelt es sich meist nicht nur um eine Einzelaktivität, sondern oftmals um ein Element einer koordinierten Kampagne. Außerdem ist Desinformation nicht nur im nationalen Kontext, sondern auch in internationalen Konflikten von Bedeutung (Schünemann 2022, S. 32, siehe auch Abschnitt 4.5). Autokratische Staaten praktizieren hingegen intern ein restriktiveres Vorgehen gegenüber den Freiheiten im Internet und der Medienregulierung und sind weniger durch Desinformation von außen mittels KI gefährdet bzw. nutzen intern Desinformation, auch per KI, zum Machterhalt.

Noch kein massives Problem – Zukunft?

Auch wenn im Zuge der Recherche für diese Studie eine Reihe von Beispielen für politisch-motivierte Deepfakes gefunden werden konnten, kann bis dato in Europa bzw. Österreich noch nicht davon gesprochen werden, dass diese Tools bereits im großen Stil zum Einsatz kämen und schon jetzt ein massives Problem darstellten. Das hängt vermutlich mit dem frühen Stadium der technischen Entwicklung zusammen, denn Deepfake-Videos, die mit den einfach verfügbaren Tools erstellt werden, wirken vielfach noch nicht so überzeugend, wie es für Irreführung notwendig erscheint. Angesichts der sehr dynamischen Entwicklung auf dem Markt der Videogeneratoren[82] kann jedoch davon ausgegangen werden, dass sich das in naher Zukunft ändern könnte.

[82] So wurde etwa von OpenAI erst im Dezember 2024 das Tool SORA veröffentlicht, das in der Handhabung ähnlich einfach wie ChatGPT oder DallE ist. Siehe dazu Abschnitt 2.2.

Langzeitstudien zu den Auswirkungen von Deepfakes fehlen noch

Es besteht weiters die Schwierigkeit, die Auswirkung von Desinformation auf demokratische Prozesse empirisch nachzuweisen und zu interpretieren. Systematische und über längere Zeitläufe – insbesondere Deepfakes sind ein relativ neues Phänomen – durchgeführte wissenschaftliche Studien zum Einfluss von Desinformation auf den politischen Diskurs fehlen noch, daher konnte dieser bisher empirisch noch nicht zweifelsfrei nachgewiesen werden:

> *„However, measuring the macro-level effects of disinformation is obviously very difficult, as this requires an all-encompassing view, probably a greater historical distance from the events of interest, and profound knowledge of the sociocultural and political configurations of a given society."* (Schünemann 2022, S. 41)

Einerseits weiß man in konkreten Fällen zu wenig über die verantwortlichen Akteure und Quellen der Desinformation oder gar über eine mögliche koordinierte Aktivität. Die Effekte sind selbst für die Täter:innen unklar, hängen aber wiederum von der fehlenden Zuschreibung ab (Schünemann 2022, S. 36). Außerdem handelt es sich um systemische und soziale Effekte, die nicht auf Eigenschaften oder Fähigkeiten von Individuen zurückgeführt werden können. In diesem Sinne ist Information vielmehr Rohmaterial, welches erst dann einen Wert erhält, wenn man ihr in einem kollektiven Diskurs eine Bedeutung und Qualität zuschreibt. Schünemann argumentiert,

> *„it would seem more appropriate to expect successful disinformation campaigns exploiting vulnerabilities already built into the target system (here the discourse)"* (2022, S. 36)

Insofern ist der Adressat der epistemischen Angriffe durch Desinformation nicht in erster Linie das Individuum, sondern der politische Diskurs auf der Makro-Ebene.

Zunehmende Verunsicherung als Resultat von Desinformation und Deepfakes

Eine zunehmende Verunsicherung als Resultat von Desinformation und Manipulation lässt sich jedoch beobachten (Karmasin et al. 2024, S. 27) und es gibt dazu bereits erste empirische Befunde. Einer aktuellen Studie der Bertelsmann-Stiftung zufolge waren 54 % der EU-Bürger:innen innerhalb der letzten Monate unsicher, ob eine Information im Internet wahr oder falsch ist (Unzicker 2023). In einer internationalen Erhebung der Risikowahrnehmung von 1.490 Expert:innen wurde Misinformation und Desinformation als größtes Risiko der nächsten Jahre wahrgenommen und liegt damit noch vor den wahrgenommenen Risiken durch den Klimawandel (WEF 2024).

4.3 POLITISCHES MICROTARGETING

In Abschnitt 4.1 haben wir gezeigt, dass Generative KI das Potenzial hat, bestehende negative Entwicklungen in der digitalen Kommunikation zu verschärfen und zusätzlich auch spezifische neue problematische Folgen auszulösen. In Abschnitt 4.2 wurden die besonderen Risiken durch Deepfakes auf unterschiedlichen Adressaten-Ebenen betrachtet. In diesem Abschnitt 4.3 fokussieren wir im Detail auf mögliche Diskursverzerrungen durch die individuelle Exposition eventuell manipulierter, generierter Inhalten im politischen Kontext.

Politiker:innen war es bis in die jüngste Vergangenheit nur in Ausnahmefällen möglich, gezielt einzelne Personen bzw. kleine und kleinste Gruppen von Wähler:innen zu adressieren, etwa im Straßenwahlkampf, telefonisch bzw. durch Hausbesuche. Politische Botschaften wurden bislang im Wesentlichen über Mas-

senmedien (Presse, Radio, Fernsehen, Plakate) verbreitet und haben viele Leute mit denselben Inhalten erreicht. Wie bereits in Abschnitt 3.4 angedeutet, werden durch sog. Microtargeting seit einigen Jahren auch gezielt Einzelne oder besondere Bevölkerungs- oder Interessengruppen mit mehr oder weniger individualisierten Botschaften erreicht.[83] Dies kann zunächst ganz konventionell über (E-)Briefe oder Telefonate realisiert werden, wobei die dafür notwendigen Daten aus Wählerverzeichnissen und sonstigen öffentlichen Daten etc. stammen können. Viele digitale Plattformen mit nutzergenerierten Inhalten (Sammelbegriff: Social Media) ermöglichen es weiters, Inhalte als (bezahlte) Anzeigen auf der jeweiligen Plattform darzustellen und den Adressatenkreis dieser Anzeigen anhand von individuellen Persönlichkeitsmerkmalen wie Alter, Geschlecht, Wohnort oder politischen Anschauungen auszuwählen. Diese zielgerichtete Verbreitung über digitale Kanäle ist auf große Zustimmung und Weiterverbreitung der Inhalte zugeschnitten (Dreyer et al. 2021, S. 10). Dabei ist vielen Nutzer:innen oft nicht bewusst, dass es sich um derart maßgeschneiderte Werbung handelt: Im Rahmen der Digital-Skills-Austria-Studie konnten nur 12 % der Befragten nutzerspezifische Werbung erkennen und verstehen (Rauschenberger et al. 2023).

Politisches Microtargeting bereits lange im Einsatz

Dieses „data-driven campaigning" (Dommett et al. 2023) ist seit längerem in der Politik im Einsatz, freilich in sehr unterschiedlichem Ausmaß und teils nur mit groben Filtern (etwa geographischen). Kruschinski (2023) argumentiert, dass in Europa Microtargetingmethoden erst relativ überschaubar angewendet werden – wobei jedoch davon auszugehen ist, dass Entwicklung und Verbreitung kontinuierlich weitergehen werden. Bereits 2008 wurde von Kandidat Barack Obama digitales Microtargeting in den USA eingesetzt. Für Deutschland ist politisches Microtargeting ebenfalls schon länger nachweisbar (Hegelich/Serrano 2019). Bezüglich Österreich konnte beobachtet werden, dass die Parteien schon im Jahr 2017 Profiling-Techniken und Microtargeting eingesetzt haben.[84] Auch abseits von Wahlkämpfen wird Microtargeting zur politischen Beeinflussung genutzt, wie ein jüngst bekannt gewordener Fall im Zusammenhang mit der EU-Kommission zeigt.[85] Microtargeting allgemein als politische Strategie ist daher Gegenstand politik- und kommunikationswissenschaftlicher Forschung (siehe z. B. Müller-Brehm 2019; Haller/Kruschinski 2020; Dommett et al. 2023), die jedoch an der nicht zufriedenstellenden Datenlage leidet.

Cambridge-Analytica-Skandal

Im Rahmen der Brexit-Kampagne zugunsten von „Leave" sowie im US-Wahlkampf 2016 zugunsten des Kandidaten Trump wurden Millionen von Facebook-Nutzer:innendaten ohne Zustimmung für gezieltes Microtargeting missbräuchlich verwendet – was nach der durchführenden britischen Firma als Cambridge-Analytica-Skandal in die Geschichte eingegangen ist (z.B. Howard 2020). 2023 warf die Datenschutzorganisation noyb mehreren deutschen Parteien vor, während des Bundestagswahlkampfs 2021 mithilfe von Microtargeting potenziellen Wähler:innen ohne deren Einverständnis gezielt personalisierte Wahlwerbung

Datenschutzproblematik

[83] Für einen ersten Überblick siehe *parlament.gv.at/dokument/fachinfos/zukunftsthemen/008_microtargeting.pdf*.

[84] *profil.at/oesterreich/brodnig-digitaler-wahlkampf-8388750*.

[85] *netzpolitik.org/2023/chatkontrolle-eu-datenschutzbeaufragter-untersucht-microtargeting-der-eu-kommission/*.

angezeigt zu haben.[86] Auch explizite Zustimmung würde jedoch nach Ansicht von Expert:innen die Datenschutzproblematik nicht lösen.[87] Dirk Helbing berichtet weiters, dass „im letzten Jahrzehnt mehr als 60 Demokratien mit solchen Methoden unterminiert worden" seien.[88] Zwar ist die Firma Cambridge Analytica nicht mehr am Markt vertreten, es gibt jedoch Unternehmen, die ihr Erbe angetreten haben und Wählerbeeinflussung, unter anderem mit Microtargeting, zu ihrem – mitunter illegalen[89] – Geschäftsmodell gemacht haben (Cotter 2022). Auch der ausführliche Bericht des Wissenschaftlichen Dienstes des Europäischen Parlaments (EPRS) (Dumbrava 2021) nennt noch vor dem großflächigen Aufkommen Generativer KI Microtargeting als eines der Hauptrisiken für die Demokratie.

4.3.1 MICROTARGETING UND GENERATIVE KI

Generative KI hat das Potenzial, Microtargeting noch effizienter zu machen

Während Microtargeting also nichts grundsätzlich Neues darstellt, sind durch Generative KI qualitative und quantitative Veränderungen zu erwarten:

- Mehr als beim „klassischen" Microtargeting, bei dem noch vergleichsweise große Gruppen angesprochen wurden, kann KI praktisch für die individuelle Ebene spezifische Botschaften generieren oder variieren. Dies kann automatisiert und parallel für viele Individuen durchgeführt werden. Auch wenn die Aussagekraft einzelner personenbezogener Parameter, die für das Targeting herangezogen werden, in Frage gestellt wird (z. B. Kruschinski 2023 in seinen Konklusionen), dürfte der Einsatz Generativer KI eine effizienzsteigernde Wirkung haben.

Psychologische Zielgenauigkeit

- Die Zielgenauigkeit und potenzielle Effektivität wird zusätzlich durch spezielle Botschaften in entsprechend kleinen Portionen („Informationshappen") erreicht, in denen Personen basierend auf psychographischen Daten emotionaler und scheinbar persönlich angesprochen werden (Zarouali et al. 2022). Die dafür notwendige und durch KI-gestützte Auswertung verlangte Kenntnis der psychologischen Verfasstheit der Mediennutzer:innen ist seit vielen Jahren Gegenstand der F&E-Aktivitäten der Plattformen, da deren Geschäftsmodell das gezielte Marketing ist (Queck/Oppelt 2018; Zuboff 2019). Da es nicht nur im Handel, sondern auch in der Politik vielfach um Emotionen geht, ist diese gezielte Werbung auch für Kundschaft aus der Politik attraktiv. Generative KI kann hier dazu eingesetzt werden, die Botschaften im Stil angepasst und automatisch zu erstellen.

Überzeugungskraft von Videos und Fotos

- Weiters erreicht Microtargeting eine neue Qualität durch Videos und Fotos, während es bislang im Wesentlichen um die Verbreitung von Texten ging. In Zukunft können auch Deepfake-Videos und -Fotos in großer Zahl automatisiert und versehen mit individuellen Botschaften erzeugt werden (zu deren hohem Beeinflussungspotenzial siehe bereits in Abschnitt 4.2).

---

86 *datenschutz-notizen.de/wahlwerbung-mittels-microtargeting-1043907/*; siehe auch *bpb.de/themen/medien-journalismus/digitale-desinformation/290522/microtargeting-und-manipulation-von-cambridge-analytica-zur-europawahl/*.

87 Siehe Stellungnahme der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK) vom 21. Juni 2023, *datenschutzzentrum.de/uploads/dsk/23-06-21_DSK-Stellungnahme_politisches_Targeting.pdf*.

88 In: Der Standard vom 10.04.2024, *derstandard.at/story/3000000214315/computersoziologe-wir-haben-wahlmanipulation-in-grossem-stil*.

89 Vgl. etwa das „Team Jorge", *de.wikipedia.org/wiki/Team_Jorge*.

– Schließlich betrifft die Individualisierung nicht nur die Herstellung der Botschaften, sondern auch die Verbreitungswege, da die Botschaften nicht unbedingt von einem einzelnen Sender (etwa der wahlwerbenden Partei), sondern potenziell von einer Vielzahl von zu diesem Zweck angelegten (Fake-)Accounts ausgehen können („Funktionäre" der Partei). Auch dies ist mit Generativer KI in immer überzeugenderer Weise automatisierbar.

Individualisierte Verbreitungswege

Potenziale von Microtargeting für die Politik

Diese Potenziale für die Politik sind freilich noch so neu, dass unklar ist, ob sie entsprechend realisiert werden. Welche Effekte diese Kombination aus Personalisierung und Emotionalisierung tatsächlich auf politische Diskurse und politische Meinungsbildung haben, ist nicht eindeutig und nach wie vor Gegenstand der Forschung (Dumbrava 2021; Karmasin et al. 2024, S. 26). Eine großflächige Beeinflussbarkeit der Meinung von Wähler:innen kann nicht als nachgewiesen angesehen werden, es gibt Befunde in beide Richtungen (siehe bspw. Hegelich/Serrano 2019; Hackenburg/Margetts 2024; Simchon et al. 2024). Dabei ist zu berücksichtigen, dass in Zukunft wohl alle wahlwerbenden Parteien diese neuen Instrumente anwenden werden, womit die (sich teilweise gegenseitig aufhebenden) Effekte immer schwerer nachweisbar sein werden. Allerdings werden die dafür notwendigen finanziellen Ressourcen je nach Größe der Gruppierung kaum gleichmäßig verteilt sein. Weiters könnten sich für die praktische Verwirklichung Bündnisse zwischen (politisch motivierten) Plattformbetreibern, wie etwa Elon Musk, und einzelnen Politiker:innen oder Parteien formen. Aber selbst ohne breite Wirkung könnten diese Methoden das sprichwörtliche Zünglein an der Waage sein, das Wahlen entscheidet.[90] Viele Beobachter:innen äußern daher Befürchtungen, dass durch Generative KI unterstütztes Microtargeting bereits bekannte und in den vorigen Abschnitten dargestellte Folgen des digitalen Strukturwandels[91] wie die Schwächung traditioneller Kontrollinstanzen, die Manipulation der Verteilung der Inhalte und die Personalisierung der Informationswelten und Kommunikationsgewohnheiten verstärken könnten. So kommt etwa eine Forschungsgruppe der Universitäten Stanford und Columbia zu folgendem Befund:

„Matching the language or content of a message to the psychological profile of its recipient (known as „personalized persuasion") is widely considered to be one of the most effective messaging strategies. We demonstrate that the rapid advances in large language models (LLMs), like ChatGPT, could accelerate this influence by making personalized persuasion scalable." (Matz et al. 2024)

Politisches Marketing via Microtargeting

Dass Politiker:innen und Parteien jeweils auch die neuesten Technologien nutzen, um ihre Positionen zu verbreiten und möglichst viele Wähler:innen von sich zu überzeugen, ist freilich legitim. Somit lässt sich zunächst feststellen, dass Microtargeting zukünftige Wahlkämpfe wohl mitgestalten wird, nicht zuletzt,

[90] Eine Möglichkeit wäre es etwa, in umkämpften Wahlbezirken gezielt potenzielle Wähler:innen der gegnerischen Partei durch entsprechend angepasste, psychologisch unterfütterte Botschaften vom Urnengang abzuhalten (Demobilisierung), etwa indem suggeriert wird, dass die Wahl ohnehin bereits für die präferierte Partei entschieden sei. Dass Soziale Medien Einfluss auf die Wahlbeteiligung haben, wurde bereits 2012 von Facebook selbst in einem Experiment (!) nachgewiesen (Bond et al. 2012, S. 295).

[91] Bereits vor dem Aufkommen von Generativer KI gab es einerseits die Vermutung, dass Microtargeting gesellschaftliche Polarisierung verstärkt, indem es eine höhere Präzision in der Werbestrategie der politischen Kandidat:innen ermöglicht (Prummer 2020). Andererseits kann Microtargeting eine Quelle für Fehlinformationen, Negativität und eine Hyperfragmentierung von Kampagnen und der Öffentlichkeit sein (López Ortega 2022).

weil der Aufwand – etwa im Vergleich zu Straßenwahlkämpfen, wo nur relativ wenige Personen direkt erreicht werden – durch Instrumente Generativer KI immer geringer werden wird – und das im Zeitalter der sich ohnehin in den digitalen Raum verlagerten Diskurse und Informationsgewinnung. Dabei ist allerdings zu berücksichtigen, dass diese Art politischen Marketings den rechtlichen Rahmenbedingungen entsprechen muss, was etwa im Zusammenhang mit dem oben erwähnten Cambridge-Analytica-Skandal nicht der Fall war. Vor allem geht es hier um den Datenschutz, was dem automatisierten Profiling freilich enge Grenzen setzt (siehe nächster Abschnitt 4.3.2).

Problematische Aspekte

Auch wenn die Grenzen zwischen illegitimer offenkundiger Manipulation und dem legitimen strategischen Verbreiten politischer bzw. ideologischer Inhalte, die noch keine unmittelbare Beeinflussungsabsicht erkennen lassen, fließend sein können (Karmasin et al. 2024, S. 27), gibt es zumindest drei Aspekte dieser Entwicklung, die als problematisch eingestuft werden können:

Bewusste Manipulation

- Während das individualisierte Zustellen von politischen Botschaften als solches unproblematisch ist, gibt es auch organisierte, politisch motivierte Akteur:innen, „die gezielt versuchen, öffentliche Diskurse und freie Meinungsbildung zu untergraben“ (Karmasin et al. 2024, S. 27) und die die neuen Instrumente dazu nutzen, mit Desinformation (Erfundenes oder bewusst Verschwiegenes) gezielt einzelne Wähler:innen zu manipulieren. Die digitale Souveränität des Staates (siehe Abschnitt 4.7) wird damit potenziell ausgehöhlt bzw. kann der freie demokratische Diskurs auch durch interne Akteure in Gefahr gebracht werden. Simchon et al. (2024) nennen die Kombination aus Microtargeting und Generativer KI sogar eine „manipulation machine“.

Widersprüchliche zielgruppenspezifische Inhalte

- Es ist ohne Zweifel legitim, die eigenen Botschaften so effizient wie möglich an die Adressat:innen zu kommunizieren und dabei gezielt einzelne Gruppen anzusprechen und auf deren besondere Situationen einzugehen. Demgegenüber ist es als problematisch einzustufen, wenn Microtargeting auch absichtlich widersprüchliche Botschaften an unterschiedliche Zielgruppen gesendet werden (sog. Dark Advertising[92]). Auch wenn es schon bisher bisweilen zur politischen Strategie gehört, für unterschiedliche Auditorien verschiedene „Sprachen” zu sprechen und bestimmte Teile des eigenen Wahlprogramms hervorzuheben, ist es doch anders zu beurteilen, wenn die unterschiedlichen Botschaften bewusst inkonsistent bleiben. „Anstatt umfassender Kenntnis der Ziele einer Partei können an unterschiedliche Personen so völlig unterschiedliche Bilder vermittelt werden; jeweils so, wie es den persönlichen Vorlieben entspricht.“ (Hügel 2019) Es ist für die freie Meinungsbildung in einer demokratischen Gesellschaft problematisch, wenn der einen Gruppe eventuell das Gegenteil oder zumindest etwas Inkompatibles als einer anderen Gruppe

[92] Für eine Definition siehe *en.wikipedia.org/wiki/Dark_advertising*.
Das Phänomen wurde erstmals im US-Wahlkampf 2016 beschrieben, vgl. *thebureauinvestigates.com/stories/2017-05-15/the-dark-ads-election-how-are-political-parties-targeting-you-on-facebook/*. Teilweise wird der Begriff auch unscharf für Microtargeting allgemein verwendet, etwa hier: *ndr.de/fernsehen/sendungen/zapp/medienpolitik/DarkAds-Der-geheime-Wahlkampf-im-Netz,darkads102.html*. Definitionsgemäß sind Dark Ads auch in dem Sinne „dark“, also dunkel, als es keine Möglichkeit gab einzusehen, wer was bekommen hat. Facebook hatte damals angekündigt (siehe *theverge.com/2017/9/21/16346828/mark-zuckerberg-facebook-political-ads*), diese Intransparenz aufzuheben.

versprochen wird, ohne dass dies, weil es sich um keine öffentliche Kommunikation handelt, überhaupt auffallen kann.[93] Dies kann bei miteinander inkompatiblem Informationsgehalt zu schleichender Fragmentierung der demokratischen Öffentlichkeit führen (Merli 2019, S. 119f).

Politischer Spam

– Wenn sich diese spezifische Art der politischen Werbung großflächig durchsetzen sollte, dann würde es vermutlich zumindest in Vorwahlzeiten zu einer Flut an Botschaften an die einzelnen Wähler:innen führen (sog. „Infodemic"). Ähnlich Werbebotschaften von Unternehmen, Massenmails von NGOs oder verführerischen Angeboten zu Pyramidenspielen u. ä. m. entstünde möglicherweise eine neue Kategorie politischen, KI-generierten „Spams" – deren Effektivität ähnlich dem bisher bekannten Spam noch zu beobachten wäre.

4.3.2 RECHTLICHE ASPEKTE

Wie eingangs erwähnt, ist Microtargeting nichts grundsätzlich Neues, daher gibt es seit einiger Zeit auch regulative Aktivitäten in diesem Zusammenhang. Einen wichtigen Rahmen setzt einerseits das Datenschutzrecht, andererseits der neue EU-Rechtsakt über die Transparenz und das Targeting politischer Werbung (vgl. dazu die juristische Untersuchung von Kopcik 2023).

Datenschutz

Die EU-Datenschutzgrundverordnung (DSGVO) lässt Microtargeting in der EU, anders als etwa in den USA, nur unter sehr spezifischen Voraussetzungen zu: Wird die mutmaßliche politische Einstellung einer Person (eine nach Art. 9 Abs. 1 DSGVO besonders schützenswerte Datenkategorie) über Korrelationen anderer Daten bestimmt (Profiling), bedarf dies der dezidierten Zustimmung der jeweiligen Person. Da diese Zustimmung in der Praxis nicht eingeholt werden kann (es geht ja um individualisierte *Massen*anwendungen), wäre das Profiling nur in anonymisierter Form zulässig (über sog. „Personas"[94]) und damit automatisch weniger zielgenau sowie immer auf etwas größere Gruppen bezogen, also nicht im engeren Sinne *Micro*targeting. Kopcik (2023, S. 96) kommt zum Schluss, dass seit der Einführung der DSGVO effizientes Microtargeting ohne entsprechende Erlaubnis auf legalem Wege fast unmöglich geworden sei.

Wahlkampfteams verwenden in der Praxis Apps, die aufgrund kumulierter Daten die Wahrscheinlichkeit für eine bestimmte Wahlentscheidung anzeigen.[95] Damit wird ein direkter Personenbezug vermieden, etwa durch die Bündelung einiger Adressen zu Wohnblocks oder Straßenzügen. Es bleibt zu untersuchen, ob Recht und Praxis im datengestützten (Haustür-)Wahlkampf übereinstimmen

---

93 Der Facebook-Mutterkonzern Meta hat mittlerweile die sog. Ad Library öffentlich gemacht (*facebook.com/ads/library/*), in der es prinzipiell möglich wäre nachzuvollziehen, wer welche (bezahlten) Posts bekommen hat. Damit ist begrenzt mehr Transparenz hergestellt, doch ist zu bezweifeln, dass viele Empfänger:innen diese Datenbank bei der Informationsaufnahme durchsuchen werden (siehe dazu auch: *bpb.de/themen/medien-journalismus/soziale-medien/545989/die-digitale-professionalisierung-der-politischen-kommunikation/#node-content-title-3*).

94 *de.wikipedia.org/wiki/Persona_(Mensch-Computer-Interaktion)*.

95 Z. B. *de.wikipedia.org/wiki/CDU_connect*. Gerade das Beispiel dieser App zeigt freilich auch, wie vulnerabel diese Technik ist, denn diese App wurde 2021 gehackt, womit geschützte Daten frei verfügbar wurden.

und insbesondere, wie sich das bei KI-gestütztem, individualisiertem Microtargeting darstellen würde.[96]

EU: Gesetz über Transparenz und Targeting politischer Werbung

Die EU hat sich des Themas (Micro-)Targeting als Spezifizierung der Regeln des Digital Services Act (DSA) angenommen: Die Verordnung (EU) 2024/900 über die Transparenz und das Targeting politischer Werbung[97] wurde im März 2024 verabschiedet und wird ab Oktober 2025 anwendbar sein. Dieser Rechtsakt ist wie viele der jüngeren Rechtsakte der EU bemerkenswert umfangreich und detailliert: er ist im Amtsblatt 44 Seiten lang, die Hälfte davon sind die 114 ausführlichen Erwägungsgründe, die den thematischen Hintergrund umfassend darstellen, die andere Hälfte die 30 substanziellen Artikel. Das dokumentiert, dass sich die EU mit diesem Thema offensichtlich intensiv befasst hat. Jurist:innen und Politikwissenschaftler:innen sind derzeit dabei aufzuarbeiten, ob der Rechtsakt tatsächlich etwas in der juristischen und politischen Praxis ändern wird.

Für das hier behandelte Thema ist die Verordnung jedenfalls einschlägig. In Art. 3, Punkt 11 des Rechtsakts werden „Targetingverfahren" definiert als „Verfahren, die eingesetzt werden, um, auf der Grundlage der Verarbeitung personenbezogener Daten, eine politische Anzeige nur an eine bestimmte Person oder Personengruppe zu richten oder diese auszuschließen". Kapitel III (insb. Art. 18 und 19) sind speziell dem „Targeting und Anzeigenschaltung politischer Werbung über das Internet" gewidmet: Im Wesentlichen braucht es auch nach diesem Gesetz die ausdrückliche Einwilligung der Beworbenen, außer es handelt sich um ehemalige oder gegenwärtige Mitglieder einer Partei oder Abonnent:innen; weiters gibt es ein Verbot der Werbung an Jugendliche (bis ein Jahr vor Erreichung des Wahlalters). Darüber hinaus müssen die mit Targeting verfolgte Vorgangsweise transparent dargestellt, Protokolle geführt, jede solche Botschaft als „politische Anzeige" gekennzeichnet sowie die Nutzung von KI explizit ausgewiesen werden. Der Rechtsakt wird erst ab Herbst 2025 angewendet werden, daher kann seine Effektivität und konkrete Umsetzung noch nicht beurteilt werden.

4.3.3 ZWISCHENFAZIT

Die meisten Beobachter:innen und sogar der europäische Gesetzgeber kommen zum Schluss, dass die Effekte von Microtargeting aus ethischer und demokratiepolitischer Sicht problematisch sind. Der Soziologe Dirk Helbing formuliert es so:

> *„Selbst wenn Pseudonyme verwendet werden, was der Datenschutz verlangt, ist es trotzdem möglich, anhand von Eigenschaften und Vorlieben der Personen Microtargeting zu betreiben und sie so zu manipulieren. Das heißt, auch die freie Wahl ist infrage gestellt. Das macht es letztlich auch möglich, dass diejenigen, die am meisten Geld investieren in die Meinungsbildung, fast so etwas wie Stimmenkauf betreiben*

---

[96] Es gibt in diesem Zusammenhang einige rechtliche Fragen, die hier nicht beantwortet werden können (vgl. auch Kopcik 2023). Zu prüfen ist etwa, ob das Verbot der Zusendung unerwünschter Sendungen laut § 107 TKG 2003 individualisierte, nicht im gewerblichen Verkehr von Privaten (bzw. politischen Akteur:innen) verschickte Nachrichten überhaupt umfasst. Aufgrund der neuartigen Individualisierung fiele solcher Spam wohl nicht unter den Begriff der (zustimmungspflichtigen) *Massen*sendung.

[97] Amtsblatt der EU L 2024/900, 20.3.2024, *eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202400900*.

können. Es gilt dann nicht mehr „eine Person, eine Stimme", sondern: Wer mehr Geld investiert, bekommt mehr Stimmen. So war die Demokratie nicht gedacht."[98]

Zwar gab es auch in Österreich in früheren Wahlkämpfen Umfrage- und Inseratendeals mit teils unerlaubter Überziehung der Wahlkampfbudgets, Helbing verweist hier somit auf die höhere manipulative Effizienz des Microtargeting im Vergleich zu früheren Vorgangsweisen, da so vermutlich mit weniger Geld mehr Impact erreicht werden kann.

Der 4. Erwägungsgrund der in Abschnitt 4.3.2 erörterten EU-Verordnung formuliert nüchterner, aber nicht weniger eindringlich:

„Die Notwendigkeit, Transparenz zu gewährleisten, ist ein legitimes Ziel des Allgemeininteresses im Einklang mit den gemeinsamen Werten der Union und der Mitgliedstaaten gemäß Artikel 2 [des EU-Vertrags][99]*. Es ist nicht immer einfach für Bürger, politische Anzeigen zu erkennen und ihre demokratischen Rechte in informierter Weise auszuüben. Die Zunahme der Komplexität der Desinformation, die Diversifizierung der Akteure, die rasche Entwicklung neuer Technologien und die verstärkte Verbreitung von Informationsmanipulation und Einflussnahme auf die demokratischen Wahl- und Regulierungsprozesse stellen für die Union und die Mitgliedstaaten wichtige Herausforderungen dar. Politische Werbung kann ein Vektor für Desinformation sein, insbesondere wenn der politische Charakter nicht aus der Werbung hervorgeht, wenn sie von Sponsoren außerhalb der Union stammt oder wenn dabei Targeting- oder Anzeigenschaltungsverfahren zum Einsatz kommen. […]"*

4.4 MACHTKONZENTRATION IM BEREICH GENERATIVE KI

Machtkonzentration im digitalen Markt schon seit langem

Die Diskussion von Machtkonzentration im Bereich der digitalen Technologien und der hiermit verbundenen Firmen ist nichts Neues. Schon 1999 hat Schiller in einer technikgeschichtlichen Arbeit herausgestellt, dass die Verbreitung von mittlerweile ubiquitären digitalen Technologien nicht zu einer Demokratisierung von Macht, sondern vielmehr zu einer Konzentration bei wenigen führt. Diese wiederum ist häufig in der Hand jener Firmen, welche die digitale Infrastruktur bereitstellen, etwa der digitalen Plattformen (Schiller 1999; Schiller 2014).

Die wenigen zentralen Unternehmen, die den digitalen Raum wirtschaftlich und zum Teil auch organisatorisch beherrschen, wie die US-amerikanischen Firmen Alphabet, Amazon, Apple, Meta und Microsoft sowie die chinesischen Firmen Tencent und Alibaba, nehmen aufgrund ihrer Größe eine besondere Stellung ein (Khanal et al. 2024).

Netzwerk- und Skaleneffekte

Zurückzuführen ist diese Dominanz auf zwei zentrale ökonomische Fak-toren: (1) Angebots- und nachfrageseitige Netzwerkeffekte und (2) Skaleneffekte, da zumindest bisher die Grenzkosten i. d. R. gegen Null gingen. So handelt es sich vielmehr um eine künstliche Verknappung eigentlich oft vorhandener Güter (Staab 2019). Gerade diese künstliche Verknappung ist schon von Schumpe-

---

98 Interview in Der Standard unter der Überschrift: „Wir haben Wahlmanipulationen in großem Stil" (10.04.2024), ***derstandard.at/story/3000000214315/computersoziologe-wir-haben-wahlmanipulation-in-grossem-stil***.

99 Dort ist festgehalten, dass die EU und ihre Mitgliedstaaten Demokratien sind.

ter als zentraler Trend zur Monopolisierung beschrieben worden (Schumpeter 1942[2013], S. 87-106). Getrieben und getragen von privatem Risikokapital und den damit einhergehenden (Monopol bzw. Oligopol-) Renditeerwartungen streben die Firmen zu einer Dominanzposition hin. All dies führt zu einem sog. Lock-In, d. h. dass die dadurch entstandenen Verhältnisse nicht mehr bzw. nur unter besonders hohem Aufwand verändert werden können.

Neue Entwicklungen

Während sich die grundsätzliche Machtkonzentration bei einigen wenigen Digitalkonzernen durch die Verbreitung von Generativer KI nicht plötzlich auflöst, gibt es jedoch erste Anzeichen dafür, dass neue Akteure (nicht-staatliche und staatliche) auftreten und somit im Gesamtsystem Veränderungen (weitere Oligopolisierung) zu erwarten sind.

4.4.1 ALTE UND NEUE FIRMEN IM KI-MARKT

Die Anzahl an Firmen, welche zumindest in der Selbstbeschreibung etwas mit KI im Allgemeinen oder konkret mit Generativer KI zu tun haben, ist derzeit aufgrund der sich ständig verändernden und dynamischen Bedingungen weder in ihrer Gesamtheit zu erfassen noch wären diese Angaben aufgrund der Geheimhaltung der Firmen in irgendeiner Form überprüfbar. Dennoch lassen sich die Firmen in drei zentrale Kategorien unterteilen (Widder et al. 2023): Jene, die Grundlagenmodelle herstellen; jene, die Infrastruktur bereitstellen; und jene, welche wiederum Anwendungen auf den Grundlagenmodellen bauen und für spezifische Anwendungskontexte adaptieren. In der Folge wird ein kurzer Überblick über die bestehenden Firmen gegeben, anhand deren etwaige Machtkonzentration verdeutlicht werden kann.

Firmen und Organisationen hinter bekannten Grundlagenmodellen

Es gibt eine Vielzahl von Modellen für Generative KI, die von verschiedenen Organisationen entwickelt werden (Widder et al. 2023; HuggingFace 2024). *OpenAI*, das sich selbst als Forschungsinstitut bezeichnet, verfolgt trotz seiner wissenschaftlichen Ausrichtung starke ökonomische Interessen und hat eine komplexe Organisationsstruktur. Bekannt ist es für Modelle wie GPT-4. *Alphabet (Google)* entwickelt derzeit die Modelle Gemini und Gemma, die die Sprachverarbeitung und -generierung weiter verbessern sollen. *Anthropic* hat das Modell Claude entwickelt, das auf sichere und zuverlässige KI-Nutzung abzielt, welches seit September 20+23 mit *Amazon* assoziiert ist (Coldewey 2024). *Meta (Facebook)* hat mit LlaMA-2 ein weiteres Modell hervorgebracht, das sowohl in der Forschung als auch in kommerziellen Anwendungen genutzt wird. Das französische Unternehmen *Mistral* ist ebenfalls ein Akteur in diesem Bereich. Die Falcon-Modelle des *Technology Innovation Institute* der Vereinigten Arabischen Emirate sind ein weiteres Beispiel. Auch das als Spin-off aus deutschen LMU München stammende *Stability AI* mit seinem Bilderzeugungsmodell Stable Diffusion ist erwähnenswert. Dieser Einblick in die Organisationen hinter Generativer KI verdeutlicht, dass es eine Vielzahl an Modellen gibt, die unterschiedlicher Natur sind und in der Praxis Anwendung finden. Dennoch ist die Anzahl an Organisationen, die tatsächlich originäre Modelle erschaffen im Vergleich zur Anzahl an Firmen, die derzeit laut Eigenaussage im Bereich Generative KI tätig sind, äußerst überschaubar.

Weitere Akteure

Erwähnenswert ist in diesem Kontext noch die französisch-amerikanische Firma *HuggingFace*, welche sich über die Entwicklung eigner Modelle hinaus als Open-Source-Plattform für unterschiedliche Generative KI-Modelle etabliert hat, auf welcher i. d. R. alle großen Grundlagenmodelle bereitgestellt werden.[100] Auch *EleutherAI* ist als Organisation, welche sich als nicht profitorientierte Forschungsorganisation versteht, die Generative KI-Modelle für Forschung und allgemeine Öffentlichkeit bereitstellt, ebenfalls zu erwähnen. Viele der genannten Modelle sind auf Infrastruktur des Anbieters *Amazon Web Services* (AWS) trainiert worden (Widder et al. 2023).

Infrastrukturunternehmen: insb. Chiphersteller und Rechenzentren

Als Infrastruktur werden Unternehmen erfasst, welche die relevante Hardware für die Erstellung und Betreibung von Generativen KI-Modellen bereitstellen. Besonders hervorzugeheben ist hier der Chip-Hersteller *NVidia* sowie der Hersteller von Spezialmaschinen für die Chipherstellung *ASME* (Widder et al. 2023), welche seit der öffentlichen Wahrnehmung von Generativer KI durch die Popularität von ChatGPT eine Vervielfachung ihres Unternehmenswertes erfahren haben. So hat *NVidia* in den letzten fünf Jahren ein Wertwachstum von ca. 2.500 % erfahren. Ebenfalls darunter erfasst werden können die Betreiber jener Rechenzentren, die für das Training und die Benutzung von Generativer KI-Modellen bereitstellen (z. B. *Amazons* AWS). Selbst wenn diese keine Dienste auf Basis von Generativer KI anbieten, so profitieren sie dennoch über die Vermietung von Rechenkapazitäten an der Popularität und dem Rechenhunger dieser neuen Technologie. Beispielsweise nutzt OpenAI exklusiv Microsofts Azure Rechenplattform im Gegenzug für milliardenschwere Investments.[101]

Adaptierungen und Dritt-Anwendungen:

Abhängigkeiten von großen Modellbetreibern

Eine große Anzahl von unterschiedlichen Firmen, die Generative KI-Modelle nutzen, entwickeln diese für eigene speziellere Anwendungen und Kundenbedürfnisse. In der Regel werden hierbei schon trainierte Modelle, wie z. B. jene von *HuggingFace* oder *OpenAI* verwendet. Darüber hinaus gibt es viele Anwendungen, die als Blackbox[102] bezeichnet werden müssen, weil man von außen nicht den genauen Aufbau ihrer Generativen KI-Anwendungen einsehen kann; so kann im Einzelfall nicht ausgeschlossen werden, dass diese Firmen dennoch (auch) eigene Grundlagenmodelle nutzen. Aufgrund der bisher verfügbaren Informationen kann jedoch angenommen werden, dass die meisten Generative KI-Firmen sich ebenfalls der großen, bekannten Modelle bedienen.

Lock-In-Gefahr

Grundsätzlich besteht für die Anbieter, welche sich auf die schon bestehenden Modelle größerer Anbieter setzen, jedoch die grundsätzliche Gefahr eines Lock-Ins, da ein Wechsel i. d. R. mit hohem Aufwand und Kosten verbunden ist (David 1985).

---

100 Siehe hier z. B. ***huggingface.co/models?p=1&sort=trending***.

101 ***openai.com/index/openai-and-microsoft-extend-partnership/***.

102 Diese „Blackbox" bezieht sich auf die Geschäfts- und Informationspolitik der Unternehmen und wäre prinzipiell regulativ veränderbar, während der Begriff „Blackbox" im Zusammenhang mit KI-Systemen meint, dass aus prinzipiellen Gründen nicht erklärbar ist, wie und warum sich ein Modell so oder anders verhält (vgl. Udrea et al. 2022).

4.4.2 OLIGO- UND MONOPOLISIERUNGSGRÜNDE

Techno-ökonomische Gründe

Es gibt eine Reihe von techno-ökonomischen Gründen, welche eine Tendenz zu noch größerer Machtkonzentration bei Digitalunternehmen durch das Aufkommen von Generativer KI nahelegen. Dies ist insbesondere vor dem Hintergrund der häufig als offen oder sogar als Open Source deklarierten Modellen für Generative KI besonders erklärungsbedürftig, da bei Quelloffenheit i. d. R. zunächst davon ausgegangen wird, dass diese Machtkonzentrationen entgegenwirkt. Basierend auf Widder et al. (2023) werden fünf zentrale Dimensionen diskutiert (1. Development Frameworks, 2. Rechenleistung, 3. Daten, 4. Arbeit und Wissen sowie 5. KI-Modelle).

1. Entwicklungsumgebungen (Development Frameworks)

Entwicklungsumgebungen – auch Coding oder Development Frameworks genannt – werden in der Softwareentwicklung eingesetzt, um die Entwicklungen zu unterstützen, zu standardisieren und zwischen Entwicklungsgruppen abzustimmen (Widder et al. 2023). Im KI-Bereich haben sich hierbei zwei Frameworks herauskristallisiert, welche von den Firmen *Meta* und *Alphabet* (indirekt) betrieben werden. Hierbei ist PyTorch *Meta* (und seit 2022 der *Linux Foundation*) zuzuordnen (Paszke et al. 2019), während TensorFlow *Alphabet* (Google) zuzuordnen ist (Abadi et al. 2015).[103]

Für die beiden Firmen entsteht hieraus ein entscheidender Vorteil. So kann alles, was von Entwicklergruppen oder der Wissenschaft mit diesen Frameworks entwickelt wird, ohne große Umstände in die eigenen Produkte von Meta bzw. Alphabet integriert werden. Dies bedeutet, dass z. B. wissenschaftliche Forschung, die mithilfe dieser Frameworks entsteht, auf schnellem Wege von diesen Firmen kommerziell verwertet werden kann. Darüber hinaus verlangen die spezialisierten Hardware-Chips (i. d. R. Graphikkarten, d. h. GPU), welche für die Erstellung von Generativer KI benötigt werden, die Anwendung von proprietärer Software, welche die Dominanz der entsprechenden Chiphersteller wie *NVidia* sichert.

2. Rechenleistung

Ein weiterer zentraler Faktor ist die Rechenleistung. So braucht es massive Rechenkraft, um Generative KI-Modelle zu betreiben. Die ökonomischen Kosten hierfür (welche sich aus Hardware, Energie und Expertise ergeben[104]) sind i. d. R. so hoch, dass selbst vermeintlich offene KI-Modelle nicht nachgebaut werden können. Hierbei wird auch deutlich, dass wir es nicht mehr mit einer Situation von, ökonomisch gesprochen, gegen Null gehenden Grenzkosten zu tun haben. So sind die Kosten, die aufgebracht wurden, um das zugrundeliegende GPT-4-Modell zu trainieren, niedriger als die wöchentlichen Kosten, um das darauf aufbauende ChatGPT zu betreiben (Patel 2023; Widder et al. 2023).

3. Daten

Das Trainieren und Betreiben von Generativer KI braucht viele Daten in guter Qualität, was zu einer immer größeren Herausforderung für die Entwicklung von Generativer KI wird. Während es eine Vielzahl an öffentlich zugänglichen Datensätzen gibt, ist die Frage des intellektuellen Eigentums hingegen häufig ungeklärt. Die Problematik wird durch die aktuelle gerichtsanhängige Klage der Zeitung New York Times gegen das Unternehmen OpenAI deutlich (Widder et al. 2023; Khanal et al. 2024).

---

[103] Siehe auch: *ai.meta.com/blog/pytorch-foundation/*.

[104] Für ein Rechenbeispiel siehe Abschnitt 2.2.

4. Arbeit und Wissen

Trotz des Wortteils „künstlich“ im Namen benötigt das Betreiben und Trainieren von Generativer KI generell sehr viel menschliche Arbeit. Hierbei ist das bestärkende Lernen durch menschliche Rückkopplung (Reinforcement Learning from Human Feedback, RLHF) zentral. Die Arbeit lässt sich in vier zentrale Schritte, Datenbeschriftung, Klassifizierung, Modellkalibrierung und Moderation unterteilen (Widder et al. 2023):

Datenbeschriftung und Klassifizierung

Zunächst findet eine Datenbeschriftung- und Klassifizierung der vorhandenen Datensätze statt. Die Generative KI soll in diesem Schritt darauf trainiert werden, für die Endkonsument:innen unerwünschte Inhalte herauszufiltern. Um dieses Ergebnis zu erreichen, bedarf es einer Einordnung der Datensätze, welche durch menschliche Arbeit vorgenommen werden muss. Diese findet durch sog. Clickworkers beispielsweise in Ländern wie Kenia statt, wobei die emotional belastendenden, ausbeutenden Arbeitsbedingungen erwähnt werden sollten.[105]

Modellkalibrierung und Moderation

Die Datensätze müssen in Folge immer wieder im Rahmen einer Modellkalibrierung und Moderation neu überarbeitet werden, was wiederum menschlicher Arbeit bedarf. Menschliche Arbeit ist nicht auf die emotionale[106] Arbeit bei der Erstellung und Moderation von Datensätzen beschränkt, sondern natürlich auch hinsichtlich Ingenieursaufgaben wie der Produktentwicklung und -wartung notwendig.

5. KI-Modelle

Ein wesentlicher Aspekt, den es zu berücksichtigen gilt, ist der Vorsprung, den einige Unternehmen (in unterschiedlicher Weise) im Bereich der KI-Modelle bereits erzielt haben. Diese Firmen haben nicht nur erhebliche Fortschritte gemacht, sondern sichern sich auch entscheidende Wettbewerbsvorteile, indem sie den Code und die Trainingsdaten ihrer Modelle nur unzureichend offenlegen. Durch diese mangelnde Transparenz haben sie einen deutlichen Vorsprung gegenüber potenziellen Nachahmern, die ohne Zugang zu diesen wichtigen Ressourcen kaum in der Lage sind, vergleichbare Modelle zu entwickeln. Es erscheint plausibel, dass diese Praxis den Wettbewerb so stark hemmt, dass die Innovationskraft im Bereich der Generative KI erheblich beeinflusst wird und es potenziell zu einer Ungleichverteilung von Marktchancen und somit zu einer verstärkten Marktkonzentration führen kann. Potenzielle Wettbewerber werden somit von vorneherein abgeschreckt.

4.4.3 MACHTKONZENTRATION UND STAATLICHE AKTEURE

Dominanz von IKT-Unternehmen auch über staatliche Akteure

Bisher galt, dass der IKT-Sektor stark durch das Zusammenspiel staatlicher und privater Akteure geprägt ist insbesondere durch die Verflechtungen der Schlüsselunternehmen der Digitalisierung mit nationalen Sicherheitsbehörden (Mazzucato 2018). Diese Annahme musste jedoch in letzter Zeit für neuere IKT-Unternehmen hinterfragt werden. So hat der Staat in den führenden Branchen der Digitalisierung weitgehend die wirtschaftspolitische Kontrolle verloren (Staab 2019). Diese Entwicklung wurde weiter durch Faktoren wie US-Risikokapital in chinesischen Firmen und Staatsbeteiligungen verkompliziert, die das traditionelle Verständnis von staatlicher Steuerung und wirtschaftlicher Autonomie in Frage stellen (Janeway 2012; Mazzucato 2018; Staab 2019, S. 20).

---

[105] Siehe dazu im Überblick Kapitel 5 zu den sonstigen, nicht direkt Demokratie-bezogenen Folgen von Generativer KI.

[106] Weil Trainieren konkret bedeutet, dass teils unangenehme, emotional aufregende Inhalte gelesen, angesehen oder angehört werden müssen (Hass, Gewalt etc.).

Ein zentraler Einflussvektor in dieser neuen Entwicklung ist die intensive Mitwirkung großer IKT-Unternehmen an der Formulierung und Regulierung neuerer KI-Technologien. So beeinflusst „Big Tech" die Politikgestaltung, besonders im Bereich der Generativen KI (Khanal et al. 2024).

Wirtschaftliche und staatliche Akteure

Wirtschaftliche Akteure dominieren mittlerweile staatliche Akteure im Bereich IKT. Diese Dominanz bedeutet jedoch nicht, dass diese Unternehmen nicht bereit wären, sich für verschiedene staatliche Zwecke einsetzen zu lassen. Tatsächlich können Behörden Anwendungen Generativer KI, wie etwa ChatGPT, für eine Vielzahl von staatlichen Zielen nutzen, einschließlich sicherheitsrelevanter oder geopolitischer Aufgaben.

Entwicklung staatlicher Grundlagen-Modelle

Gleichzeitig ist es durchaus denkbar, dass staatliche Akteure, die über die notwendigen Ressourcen verfügen, selbst in den Bereich der KI-Entwicklung einsteigen. Aus geopolitischen Gründen könnten sie eigene Grundmodelle entwickeln und betreiben, um ihre strategischen Interessen zu wahren und ihre Position auf der globalen Bühne zu stärken.[107] Diese Doppelrolle, bei der wirtschaftliche Akteure sowohl eigene Ziele verfolgen als auch staatliche Interessen unterstützen, sowie die aktive Beteiligung staatlicher Akteure an der Entwicklung und dem Betrieb von KI-Modellen, könnte zu einer komplexen Verzahnung von wirtschaftlichen und staatlichen Interessen führen.

Militärische Machtkonzentration

Generative KI wird zunehmend von staatlichen Akteuren auch im militärischen Bereich eingesetzt. Ein prominentes Beispiel hierfür ist die Nutzung dieser Technologie durch das israelische Militär im Rahmen von Project Nimbus, bei dem Alphabet (Google) die Cloud-Infrastruktur bereitstellt. Nimbus zielt darauf ab, KI-Technologien in militärischen Anwendungen zu integrieren, um die Effizienz und Effektivität der Streitkräfte zu erhöhen (Grant 2024). Ähnliche Entwicklungen sind auch in anderen Ländern zu beobachten. So setzt Russland Generative KI in Form von Bots ein, die in verschiedenen Cyberoperationen und Informationskriegen genutzt werden.[108]

Informationsweitergabe durch Generative KI

Auch die USA und die NATO treiben den Einsatz von Generativer KI im Kriegseinsatz voran[109] (Oniani et al. 2023). Diese Technologien werden unter anderem zur schnellen Informationsaufbereitung und -weitergabe eingesetzt, wodurch eine neue Dimension der Kriegsführung entsteht. Die zunehmende Nutzung von Generativer KI im militärischen Kontext durch verschiedene Staaten kann zu einer Machtkonzentration führen, da die Möglichkeit besteht, dass diese Technologien strategische Vorteile bieten.

---

[107] Dazu siehe auch Abschnitt 4.7 zur Digitalen Souveränität.

[108] Siehe dazu die Abschnitte 4.1, 4.2 und 4.6.

[109] The North Atlantic Treaty Organization, Summary of the NATO artificial intelligence strategy, *nato.int/cps/en/natohq/official_texts_187617.htm* (2021).

4.4.4 ZWISCHENFAZIT

Neue Oligopole

Somit können wir als Zwischenfazit die Machtkonzentration im Bereich Generative KI wie folgt ökonomisch einordnen: Die Generative KI und der damit verbundene Aufwand, also konstante oder sogar steigende Grenzkosten bei der benötigten Rechenleistung[110] in Kombination mit neuen Netzwerkeffekten (durch regelbasierte Entwicklungsumgebungen, KI-Modelle, Arbeit) führen paradoxerweise zu einer neuen Form von Knappheit. Dies begünstigt die Entstehung von faktischen 'natürlichen Monopolen' in der Technologiebranche (Posner 1978; Narechania 2021).

Gesteigerte Machtkonzentration der Digitalkonzerne

Digitalkonzerne in neuen gesellschaftlichen Sphären

Während digitale Konzerne schon lange in unterschiedliche Bereiche des gesellschaftlichen Lebens (Sphären) eindringen, treibt Generative KI diesen Prozess nun schneller und fundamentaler voran (EGE 2023, S. 27ff). Die großen Digitalkonzerne sind mittlerweile in alle Lebens- und Gesellschaftsbereiche involviert. So mag es auch nicht verwundern, dass die meisten diese Konzerne etwa im Gesundheitsbereich mit Generativer KI vertreten sind. Diese Anwendungen verfolgen das Ziel, die ärztliche Anamnese und ärztliche Dokumentation zu unterstützen.[111] Vergleichbare Beispiele können in sehr unterschiedlichen Sektoren wie Politik, Nachrichten und Medien, Recht und Wissenschaft gefunden werden.[112]

Die Rolle staatlicher Akteure ist im Kontext der generativen KI noch unklarer. Auf der einen Seite zeigt sich eine verstärkte Abhängigkeit von den etablierten Digitalkonzernen aufgrund des hohen Aufwands und der Kosten für die Entwicklung und den Betrieb eigener Grundlagenmodelle. Auf der anderen Seite können und werden staatliche Akteure als Kunden ebenjener Digitalkonzerne Funktionalitäten nutzen, sowohl für die Veränderung bürokratischer Prozesse als auch für militärische und geheimdienstliche Tätigkeiten. Staaten, die den Aufwand und die erheblichen Kosten schultern können, werden mutmaßlich aus geostrategischen Gründen selbst Grundlagenmodelle entwickeln und anwenden.

4.5 HYBRIDE BEDROHUNGEN

Hybride Bedrohung

Ein derzeit breit diskutiertes Phänomen, zu dem die Nutzung von (Generativer) KI beitragen und das auch zur Gefährdung demokratischer Prozesse führen kann, wird unter dem Titel Hybride Bedrohungen zusammengefasst. Aus dem militärischen Kontext kommend, werden unter diesem Begriff heute all jene Gefahren erfasst, welche nicht direkt auf klassische militärische Maßnahmen setzen, jedoch im Krieg oder zu dessen Vorbereitung sowie der gegnerischen Abschreckung eingesetzt werden können. Ziel ist es beispielsweise, die öffentliche Meinung in der gegnerischen Bevölkerung durch Desinformation zu beeinflussen (Villar García

Desinformation

---

110 Hier sei auf die Aussage des OpenAI-CEO Sam Altman von 2023 verwiesen, der hoffe, dass ChatGPT derweil keine neuen Nutzer:innen bekomme, da diese die Kosten für den Betrieb erhöhen würden. An dieser Aussage sieht man, dass Generative KI nicht zu abnehmenden Grenzkosten führt. Vielmehr steigen die variablen Kosten mit jeder Nutzung (Oremus 2023).

111 *healthcaredive.com/news/amazon-Generative-ai-clinical-documentation-healthscribe/688996/*.

112 *sphere-transgression-watch.org*.

et al. 2021, S. 1), um damit die öffentliche Ordnung zu destabilisieren, Kommunikationsinfrastrukturen anzugreifen und demokratische Strukturen zu schwächen und somit auch militärisch einen Vorteil zu haben (Mazzucchi 2022). Zur hybriden Kriegsführung zählen auch Sabotageakte auf fremdem Territorium. Die Terminologie von Hybriden Bedrohungen ist stark von der NATO geprägt. Die NATO und die EU haben sogar das European Centre of Excellence for Countering Hybrid Threats gegründet, bei dem auch Österreich Mitglied ist. Dieses Zentrum hat ein spezielles Toolkit zu diesem Thema veröffentlicht (Heap et al. 2021).

Hybride Bedrohung als analytisches Konzept oder politisches Programm

Es erscheint jedoch ratsam, kritisch mit dem Begriff der hybriden Kriegsführung oder Bedrohung umzugehen, da dieser eher als politisches Programm denn als analytisches Konzept anzusehen ist (Libiseller 2023). Der Begriff wird derzeit in Europa hauptsächlich im Kontext Russlands und Chinas verwendet und bezieht sich oft auf Cyberattacken und Desinformationskampagnen.[113]

Dual use

Deepfake-Technologien zählen somit zu den Dual-Use-Technologien, die nicht nur für zivile (z. B. in der Filmindustrie), sondern auch für militärische Zwecke genutzt werden können (Byman et al. 2023). Hierbei ergeben sich zwei Hauptanwendungsfelder: die Avatar-Erstellung und das Phänomen der gefälschten Sprache.

Avatar-Erstellung

Unter Avatar-Erstellung verstehen wir in diesem Kontext die günstige und vielfältige Erstellung von künstlichen Nutzerprofilen auf z. B. Social-Media-Plattformen, um dadurch interpersonelle Netzwerke von Zielpersonen oder Zielbevölkerungen zu beobachten, sich in politische Debatten und Diskussionen einmischen zu können oder um Phishing[114] zu betreiben. Durch solche fiktive Avatare, die als Lautsprecher fungieren, können Angreifer mithilfe des sogenannten „fictitious algorithmic projecting", welches auf massenhafter Erstellung von künstlichen Profilen beruht, den Eindruck eines Konsensus in einer Gruppe erzeugen (Mazzucchi 2022). Ein Beispiel hierfür ist der Versuch, die allgemeine Wahrnehmung russisch kontrollierter Gruppen in Mali und Zentralafrika zu bestärken, um die Wagner-Gruppen dort zu unterstützen.

Spamouflage

Eine Methode ist die der sogenannten „Spamouflage" [115], bei der legitime Sorgen der Bevölkerung oder soziale Konfliktthemen (wie z. B. Migration und Gender) verwendet werden, um sie durch ficticious alorithmic projecting größer erscheinen zu lassen, als sie möglicherweise in einer Bevölkerung sind. Dieser Eingriff in demokratische Willensbildungsprozesse kann auch Teil einer hybriden Strategie sein, um die Zustimmung oder Ablehnung zu Politiken oder kriegerischen Auseinandersetzungen zu beeinflussen (Matasick et al. 2024).

Weitere Beispiele

In diesem Kontext gibt es z. B. Meldungen von Versuchen russischer Akteure, während der EU-Wahl einzugreifen. So wurden laut dem Cybersecurity-Unternehmen Mandiant von der russischen Hackergruppe „Sandworm" erbeutete sensible Informationen zur Störung der Wahlauseinandersetzung weiterverbreitet. Hierbei können Massen von künstlichen Avataren in sozialen Netzwerken den nötigen Multiplikatoreffekt erzeugen.[116] Auch Artikel von propagandistischen Nachrichten-Seiten, wie die von Moskau finanzierte Nachrichtenseite Voice of

---

113 Vgl. *bmeia.gv.at/themen/globale-themen/hybride-bedrohungen*.

114 Unter „Phishing" versteht man Versuche, sich über gefälschte Webseiten, E-Mails etc. als vertrauenswürdige:r Kommunikationspartner:in in einer elektronischen Kommunikation auszugeben, um damit an persönliche Daten zu gelangen (siehe auch Abschnitt 4.6).

115 Profil 20.4.24, Die Schlacht vom 9. Juni, von Elena Crisan.

116 *orf.at/stories/3354789/*.

Europe, die verzerrte und falsche Berichte erzeugte, um in der EU-Stimmung gegen den Krieg in der Ukraine zu machen, konnte so weiterverbreitet werden[117] (Cabinet Office 2018, S. 52).

Gefälschte Sprache

Eine weitere Anwendung von Generativer KI im Kontext hybrider Bedrohungen sind die Versuche, mit gefälschten Bildern, Videos oder Sprache von z. B. gegnerischen Politiker:innen oder Persönlichkeiten die gegnerische Bevölkerung oder Armee zu verwirren, einzuschüchtern und so in demokratische Willensbildungsprozesse einzugreifen, um die Zustimmung zur Politik jener Personen zu verringern. Es handelt sich somit um gefälschte Sprache sozusagen im echten System. In diese Kategorie fällt das von Russland erstellte Video des ukrainischen Präsidenten Wolodymyr Selensky, der seine Landsleute nach dem Einmarsch des russischen Militärs im Februar 2022 vermeintlich zur Kapitulation auffordert. Es ist denkbar, dass völlig gefälschte Ereignisse oder Personen künftig die internationalen Nachrichten überschwemmen. Der kriminelle/militärische Missbrauch der Deepfake-Technologie kann demnach zur politischen Destabilisierung eingesetzt werden.

Kostenwirksamkeit für Angreifer

Während es im Jahr 2022 (Mazzucchi 2022) noch als besonders kostspielig galt, solche Informationsoperationen durchzuführen, da hierfür spezifische technische Fähigkeiten erforderlich waren, hat sich die Lage inzwischen deutlich verändert. Die einfache Verfügbarkeit von KI-Tools, die weltweit von westlichen Herstellern angeboten werden, ermöglicht bereits jetzt deren Einsatz durch praktisch jede Person. Westliche Hersteller spielen jedoch zugleich eine entscheidende Rolle bei der Bekämpfung des Missbrauchs dieser Anwendungen und ergreifen Maßnahmen, sobald sie auf solchen Missbrauch aufmerksam werden. Umfangreicher Missbrauch macht dessen Bekämpfung wahrscheinlicher. Die massenhafte Anwendung solcher Tools erfordert jedoch die Entwicklung eigener Modelle der Angreifer, um eine Abhängigkeit von westlichen Herstellern zu vermeiden. Die ökonomischen Kosten und Herausforderungen für die Erstellung eigener Grundlagenmodelle werden in Abschnitt 4.4 detailliert behandelt. Es ist zu erwarten, dass große Akteure wie Russland über die notwendigen Ressourcen verfügen, während dies für kleinere nichtstaatliche Akteure in großem Umfang unrealistisch bleibt. Sollten jedoch die Kosten für KI-Chips sinken, könnte sich die Situation in der Zukunft möglicherweise anders darstellen.

Digitale Kompetenz

Ein wirksames Mittel zur Resilienz gegenüber hybriden Angriffen ist die digitale Kompetenz der Zivilbevölkerung, die ein Bewusstsein für das potenzielle Interesse an Manipulation entwickelt (Mazzucchi 2022, siehe auch Abschnitt 6.2). Allerdings stehen diesem Ansatz unterschiedliche Interessenlagen innerhalb der Bevölkerung sowie im politischen Spektrum entgegen. Insbesondere im Kontext der aktuellen Entwicklungen in vielen EU-Staaten ist es möglich, dass sich für die meisten Themen ein politischer Akteur findet, der diese Ansichten teilt.[118] Dies bedeutet, dass trotz der generellen Anerkennung der Bedeutung digitaler Kompetenz als Schutzmaßnahme, unterschiedliche politische und gesellschaftliche Gruppierungen möglicherweise variierende Prioritäten und Meinungen hinsichtlich der Umsetzung und Förderung solcher Maßnahmen haben könnten. Diese Differenzen können die kollektive Anstrengung zur Stärkung der digitalen Resilienz erschweren.

---

117 *orf.at/stories/3352833/*; *orf.at/stories/3353041/*.

118 Siehe etwa *derstandard.at/story/3000000244540/flut-an-fakes-nach-ueberschwemmungen-in-valencia*.

Rolle der Social-Media-Plattformen

Das wichtigste Feld für diese Art hybrider Bedrohungen sind die sozialen Medien. Social-Media-Betreiber stehen freilich vor erheblichen technischen und organisatorischen Herausforderungen bei der Entdeckung von durch Generative KI erzeugten Fake-Accounts, die zur Verbreitung von Falschmeldungen oder verzerrten Informationen im Rahmen von militärischen oder nachrichtendienstlichen Methoden genutzt werden. Technisch gesehen erfordert die Identifizierung solcher Accounts den Einsatz fortschrittlicher Algorithmen und maschinellen Lernens, um Anomalien und ungewöhnliche Verhaltensmuster zu erkennen, die auf eine automatisierte Erstellung hinweisen könnten (siehe Abschnitt 2.3). Diese Algorithmen müssen ständig aktualisiert und verfeinert werden, um mit den Entwicklungen Generativer KI mithalten zu können. Eine der technischen Herausforderungen besteht darin, dass Generative KI in der Lage ist, sehr realistische und menschlich wirkende Inhalte zu erstellen, die von traditionellen Erkennungsmechanismen schwer zu erkennen sind (siehe Abschnitt 2.3). Organisatorisch gesehen müssen Social-Media Plattformen umfassende Richtlinien und Verfahren implementieren, um verdächtige Aktivitäten effizient zu überwachen und zu überprüfen. Dies erfordert nicht nur erhebliche personelle und finanzielle Ressourcen, sondern auch die Zusammenarbeit mit externen Expert:innen und Institutionen, um aktuelle Bedrohungen und Techniken zu verstehen und angemessen darauf zu reagieren. Darüber hinaus ist ein Wille zur Mitwirkung vonseiten der Plattformen nicht immer erwartbar. Plattform-Betreiber können neben ihren eigenen wirtschaftlichen Interessen, welche einer Regulierung entgegenlaufen können, weiters auch eigene politische Interessen verfolgen, welche durch den Einsatz von Generativen KI-Bots verstärkt werden können.

Dual-Use: Exportverbote greifen kaum

Ein zentrales Thema ist die Frage, welche Maßnahmen ergriffen werden können, wenn Technologien für militärische Zwecke genutzt werden, obwohl deren Verbreitung im Konsumbereich eigentlich erwünscht ist. Da die Technologien der Generativen KI nicht als (rein) militärisch klassifiziert werden (sondern als Dual-Use-Technologien), ist es nicht möglich, deren Verbreitung vollständig zu verbieten, wie es bei klassischen Waffen durch Maßnahmen zur Nichtverbreitung (Non-Proliferation) der Fall sein könnte. Stattdessen wird versucht, die Verbreitung von spezialisierten Chips, die für den Betrieb und die Entwicklung von KI-Anwendungen unerlässlich sind, einzuschränken. Ein aktuelles Beispiel hierfür ist seit 2022 die Maßnahme der USA, Herstellern wie Nvidia den Export solcher Chips in bestimmte Länder, wie beispielsweise China, zu verbieten (Shivakumar et al. 2024). Diese Exportverbote sollen verhindern, dass diese Technologien in Länder gelangen, die als potenzielle Bedrohung angesehen werden. Allerdings zeigen aktuelle Berichte, dass solche Maßnahmen häufig umgangen werden. Dies geschieht beispielsweise durch den Einsatz von Zwischenhändlern oder durch den Import von Komponenten über Drittstaaten, wodurch die Wirksamkeit der Exportkontrollen erheblich eingeschränkt wird.[119] Diese Problematik konnte bereits in den Jahren beobachtet werden, als Verschlüsselung strengen Exportkontrollen unterlag, die letztendlich jedoch nicht gegriffen haben.[120]

[119] NYTimes 05.08.2024, *nytimes.com/2024/08/04/technology/china-ai-microchips-takeaways.html.*

[120] *en.wikipedia.org/wiki/Export_of_cryptography_from_the_United_States*.

4.6 KI-CYBERKRIMINALITÄT

Die jüngsten Fortschritte Generativer KI erweitern einerseits die Möglichkeiten für kriminelle Aktivitäten, andererseits sind Generative KI-Systeme auch neuartigen Bedrohungen ausgesetzt. Viele der damit einhergehenden Gefahren sind nicht nur für demokratische Prozesse relevant, entfalten in diesem Kontext aber spezielle Wirkung – sowohl auf der Ebene einzelner Personen, als auch auf der Ebene demokratischer Abläufe und Institutionen.

Verschiedene Arten von Cybercrime

Bei Cybercrime wird grob zwischen zwei Typen unterschieden (Lallie et al. 2021), wobei die Trennung nicht ganz scharf ist: Einerseits Cyberkriminalität, die auf Informations- und Telekommunikationstechnologien angewiesen ist, wie etwa das Einbrechen in elektronische Systeme („Hacking"), das Lahmlegen von IKT-Infrastruktur (mittels sogenannter Denial-of-Service-(DoS-)Attacken) und Schadsoftware (Malware). Andererseits werden prinzipiell auch ohne IKT durchführbare Straftaten durch deren Nutzung einfacher und effektiver, wie z. B. Erpressung, Betrug und Einschüchterung, aber auch physische Einbrüche (z. B. durch das Überwinden „smarter" oder biometrischer Zugangskontrollen). Das Beispiel Datendiebstahl verdeutlicht, dass die Grenze zwischen diesen Typen nicht eindeutig ist: Einerseits können Daten auch „analog" entwendet werden, andererseits fallen durch IKT-Systeme in der Regel große Datenmengen an, die außerdem ohne physischen Zugriff entwendet werden können (und damit den Kreis der potentiellen Angreifenden und Betroffenen bedeutend vergrößert). Für all diese Bereiche bedeutet Generative KI auch neue Fähigkeiten and Angriffsmöglichkeiten für Kriminelle.

Das hat auch besondere Relevanz für Politik und Demokratie. Die Möglichkeiten, die Generative KI Cyberkriminellen bietet, um die Repräsentant:innen und Institutionen der Demokratie anzugreifen, kann dazu führen, dass Angriffe auf zentrale Pfeiler der demokratischen Gesellschaft häufiger, gezielter und damit gefährlicher werden.

4.6.1 NEUE FÄHIGKEITEN FÜR CYBERKRIMINELLE

Kriminelle nutzen Generative KI zur Imitation und zur Arbeitserleichterung

Durch Generative KI bekommen auch Kriminelle neue Werkzeuge an die Hand, mit denen sie Straftaten anders, oft effizienter und effektiver, umsetzen können. Dabei machen sie sich dieselben Eigenschaften Generativer KI zunutze, die sie auch für andere Kontexte interessant machen. Einerseits weisen aktuelle Generative KI-Systeme eine verblüffende Fähigkeit zur Imitation auf: So können gewisse Kunststile oder Individuen bei Bildern nachgeahmt werden ebenso wie menschliche Konversationen oder auch den Stil einzelner Personen oder Textformen (z. B. behördliche Dokumente). Andererseits können manche Arbeitsschritte durch Generative KI massiv erleichtert werden – so auch im Bereich Cybercrime, etwa durch die (teil-)automatisierte Erstellung von Schadsoftware und Angriffstechniken.[121]

Potenzial für Betrug, Einschüchterung und Erpressung

Wie bereits in Abschnitt 4.2.1 erwähnt, bietet Generative KI mittels verschiedener Spielarten von Deepfakes neue Möglichkeiten, Personen mittels gefälschter Bild-, Video- und Tonaufnahmen zu betrügen, unter Druck zu setzen, sie ein-

[121] Siehe bspw. diesen Bericht über Spanien: *derstandard.at/story/3000000244540/flut-an-fakes-nach-ueberschwemmungen-in-valencia*.

zuschüchtern oder zu erpressen.[122] Jüngste Fortschritte beim Klonen von Stimmen erlauben es etwa, bereits aus einer 15 Sekunden kurzen Aufnahme der Originalstimme eine überzeugende Imitation zu erstellen.[123] Entsprechend sind von Deepfakes für in der Öffentlichkeit stehende Personen (und damit einhergehendes umfangreiches Medienmaterial) eine besondere Gefahr.

Ein besonders problematischer Anwendungsbereich der Deepfake-Technologie ist die Pornografie. Bei 96 % der im Jahr 2019 online gefundenen Deepfake-Videos handelte es sich um pornografische Inhalte. Meist handelt sich hierbei um die Technik des Gesichtstausches („Face-Swapping"), bei dem der Kopf einer Person auf den Körper einer Pornodarstellerin oder eines -darstellers montiert wird. Wenn diese Videos ohne die Zustimmung der abgebildeten Person erstellt und verbreitet werden (was häufig der Fall sein dürfte), spricht man von „non-consensual pornography". Missbräuchliche Anwendungen pornographischer Inhalte gibt es insbesondere im politischen Bereich, kann aber auch im privaten Bereich bedeutsam werden. Auf der anderen Seite ist jedoch auch zu erwähnen, dass Deepfake-Pornos, die über Face-Swapping hinausgehen, sondern komplett synthetisch erzeugt werden, gerade im Bereich Kindesmissbrauch entstehen hier neue Problemfelder.[124]

Angriffe auf Politiker:innen und Institutionen der Demokratie

Eine Veröffentlichung gefälschter Inhalte kann dazu führen, dass die Betroffenen, insbesondere in der Politik, den Gegenbeweis antreten müssen, also die Fälschung nachweisen. Allein die Androhung der Veröffentlichung (oder auch Weitergabe an Vorgesetzte, Familienangehörige etc.) kann dazu genutzt werden, Betroffene unter Druck zu setzen und zu Handlungen gegen ihren Willen zu nötigen. Da Amtsträger:innen und Repräsentant:innen besonders stark in der Öffentlichkeit stehen, sind sie besonders gefährdet für derartige Angriffe. Denn häufig gibt es vergleichsweise viele Informationen über diese Personen, inklusive Bild-, Audio und Videomaterial. Damit sind sie (bzw. ihr direktes Umfeld) als Ziele von Erpressung und Diskreditierung unter Zuhilfenahme Generativer KI und Deepfakes prädestiniert.

Erleichterung von Social Engineering

Generative KI ist auch geeignet, neue Betrugsmaschen mittels sogenanntem Social Engineering umzusetzen. Dies umfasst Versuche von kriminellen Akteuren, Einzelpersonen unter dem Vorwand, dass sie mit einer legitimen Partei interagieren, dazu zu bringen, eine Handlung auszuführen (z. B. Informationen zu teilen oder eine Website zu besuchen). Solche Manipulationsversuche sind insbesondere dadurch erfolgreich, dass mit Textgeneratoren auch Sprach- und Schreibstile imitiert werden können (Ciancaglini 2020). Andererseits ist es durch das Imitieren von Stimmen per Telefon oder sogar Deepfakes in Video-Calls möglich, sogar in diesen Situationen die Identitäten anderer anzunehmen. Ein aufsehenerregender Fall ereignete sich etwa in Hongkong, wo in einer Videokonferenz mehrere (!) Manager per Deepfake nachgestellt wurden und der angebliche Chief Financial Officer eine Zahlung anwies.[125] Insgesamt verlor die Firma durch diesen Betrug umgerechnet 23 Millionen Euro. In einem anderen Fall wurden am Telefon Stimmen von Angehörigen imitiert, um vorzugeben, dass diese in einer

---

122 *onlinesicherheit.gv.at/Services/News/Audio-Deepfake-Voice-Cloning.html*.

123 NYTimes 05.08.2024, *nytimes.com/2024/08/04/technology/china-ai-microchips-takeaways.html*.

124 *tagesschau.de/investigativ/report-mainz/internet-ki-pornografie-kinder-100.html*.

125 *scmp.com/news/hong-kong/law-and-crime/article/3250851/everyone-looked-real-multinational-firms-hong-kong-office-loses-hk200-million-after-scammers-stage*.

Notsituation wären.[126] Gerade auch im politischen Kontext und im Zeitalter von Videokonferenzen können derartige Deepfakes politische Prozesse unterminieren. Generative KI ist dafür zwar nicht immer notwendig, kann jedoch bereits bislang angewandte Methoden perfektionieren.

Massen-Spear-Phishing: Vereinfachung und Individualisierung von Phishing

Eine Variante von Social Engineering sind durch Generative KI ausgefeiltere Phishing-Angriffe: Dabei wird versucht, per Email, SMS oder Messenger-Diensten die Adressat:innen dazu zu bringen, eine legitim wirkende Webseite aufzusuchen (um dort dann häufig sensitive Information einzugeben) oder einen Anhang zu öffnen (der i. d. R. Schadsoftware enthält). Herkömmliche Phishing-Angriffe zielen dabei unspezifisch auf eine große Personengruppe ab, etwa Kund:innen von Banken oder Paketdienstleistern. Durch die große Anzahl an potentiellen Opfern ist es oft ausreichend, wenn ein kleiner Anteil auf diesen Betrug hereinfällt. Dem gegenüber steht sogenanntes Spear-Phishing (Hazell 2023), das Einzelpersonen mit maßgeschneiderten Angriffen manipulieren möchte – etwa indem den Angreifern bekannte persönliche Details oder Informationen über die Organisation in den Phishing-Nachrichten verwendet werden.

Personen öffentlichen Interesses besonders gefährdet

Beide Formen von Angriffen sind durch Generative KI einfacher und in größerem Maßstab möglich: Einerseits waren Phishing-Emails oft an fehlerhafter Sprache erkennbar, andererseits ist das Erstellen von Spear-Phishing-Nachrichten zeitaufwändig. Mittels Generativer KI können einerseits qualitativ hochwertigere Massen-Phishing-Mails erstellt werden, andererseits können Informationen über Einzelpersonen automatisiert als Basis für maßgeschneiderte Nachrichten verarbeitet werden. Auch diesen Gefahren sind Politiker:innen und Amtsträger:innen besonders ausgesetzt, da über sie mehr Details öffentlich zugänglich sind als für viele andere Personengruppen. Zum Beispiel sind KI-basierte Anrufe mit imitierter Stimme oder Emails mit Verweis auf plausible Hintergrundinformationen (z. B. Reisetätigkeit, anstehende Termine) dank Generativer KI leichter zu bewerkstelligen als jemals zuvor. Ein Forscher konnte etwa mittels LLMs überzeugende Spear-Phishing-Emails erstellen, die geschickt biographische Details britischer Parlamentarier:innen einbauten (Hazell 2023).

Destabilisierung demokratischer Institutionen und Schwächung des Vertrauens

Doch nicht nur Amtsträger:innen, Repräsentant:innen und deren Umfeld können in die Irre geführt werden. Es könnten auch Bürger:innen durch Nachahmung (sog. Impersonation) behördlicher Stellen gezielt und umfassend getäuscht werden. Damit kann einerseits direkter Schaden für die Addressat:innen einhergehen, andererseits aber auch das Vertrauen in die betroffenen Institutionen – und in weiterer Folge das demokratische Gefüge insgesamt – untergraben werden.

Überlistung von Authentifizierungs-systemen

Weiters können Methoden Generativer KI auch dazu genutzt werden, um Authentifizierungssysteme zu überlisten. Bereits in Abschnitt 2.2 wurde auf die Möglichkeit des Face Morphings hingewiesen, sodass sich mit einem Dokument zwei Personen plausibel ausweisen können (Damer et al. 2018, S. 1). Das könnte etwa dazu ausgenutzt werden, um zwei Personen Zutritt zu parlamentarischen Räumlichkeiten zu ermöglichen – einer legitimerweise, einer Person hingegen ohne Berechtigung.

Programmieren von Schad-Software

Ein gänzlich anderer Anwendungsfall von Generativer KI dient dazu, Angriffe auf IKT-Systeme zu verbessern. Das kann dadurch geschehen, dass Generative KI hilft, Schadsoftware zu programmieren (Gupta et al. 2023). Das hat zweifache Auswirkungen: Einerseits werden diese Prozesse beschleunigt, andererseits

126 *washingtonpost.com/technology/2023/03/05/ai-voice-scam/*.

können auch Angreifer:innen mit wenig Wissen, anspruchsvolleren Schadcode entwickeln (Costa/Coelho 2024). Einen Vorgeschmack darauf können die Angriffe auf Partei-Webseiten kurz vor der Nationalratswahl 2024 darstellen.[127]

Generative KI zur Herstellung von Schadsoftware hat auch dazu geführt, dass Kriminelle entsprechende spezialisierte Services anbieten und den Zugang an andere Kriminelle verkaufen (Erzberger, 2023). Forscher:innen konnten sogar ChatGPT dazu bringen, zumindest ansatzweise dafür genutzt zu werden, obwohl es eigentlich derartige Anfragen verweigern sollte (Gupta et al. 2023). Dazu wird sogenanntes Jailbreaking verwendet, ein Angriff auf Generative KI, den wir im nächsten Abschnitt näher erläutern.

4.6.2 ANGRIFFE AUF GENERATIVE KI

Während in obigem Abschnitt skizziert wurde, wie sich Cyberkriminelle Generative KI zunutze machen können, fokussiert dieser Abschnitt darauf, wie Generative KI selbst Ziel von kriminellen Angriffen werden kann – und somit deren Verwendung ein Sicherheitsrisiko darstellen kann.

Leaken von (vertraulicher) Information ….

Eine große Gefahr von Generativer KI ist das Leaken von (sensitiver) Information. Dabei kann man grob zwei Datenquellen unterscheiden: Eingaben der Nutzer:innen sowie Datenbestände, die zum Trainieren verwendet wurden. Die Benutzung von Chatbots kann leicht dazu animieren, sensitive Informationen Preis zu geben, etwa um bessere Ergebnisse zu erzielen. Nach einem derartigen Vorfall verbot etwa Samsung die Verwendung von ChatGPT u. ä.[128] Daher wäre die verlässliche Wahrung des Datenschutzes wichtig, diese ist aber bei Chatbots schwer herzustellen. Oft werden etwa auch ältere Chats gespeichert oder Eingaben für das weitere Training verwendet (und somit möglicherweise in Zukunft an andere Nutzer:innen ausgegeben). Da auch Chatbots komplexe Software sind kann es sein, dass sie Sicherheitslücken haben, womit Datenlecks[129] einerseits von Kriminellen ausgenutzt werden können. Andererseits ist es möglich, dass ein Anbieter eines Generativen KI-Tools kriminelle Absichten hegt und Daten von Nutzer:innen abgreift.

… die von Benutzer:innen eingegeben wurden …

… oder in den Trainingsdaten enthalten sind

Werden von Unternehmen und Organisationen hingegen interne Dokumente, möglicherweise mit sensiblen Informationen (z. B. über Personen und Geschäftsgeheimnisse) zum Training von Generativer KI verwendet, ist es leicht möglich, dass diese Informationen an Nutzer:innen des Chatbots ausgegeben werden. Es ist auch möglich, dass Daten zwischen einzelnen Nutzer:innen unbeabsichtigt offengelegt werden (z. B. fremde Chatverläufe). Auch etwaige Vorsichtsmaßnahmen sind oft umgehbar – gerade aufgrund der hohen Komplexität von Generativer KI sind ausreichende Vorsichtsmaßnahmen nicht einfach zu treffen. Trotz dieser Herausforderungen scheint es für demokratische Institutionen und Behörden besonders wichtig, adäquate Vorkehrungen zu treffen, dass sensible Informationen nicht durch Generative KI-Anwendungen geleakt oder aktiv gestohlen werden.

Vorkehrungen gegen Datenleaks durch KI in demokratischen Institutionen

---

127 *orf.at/stories/3371205/*.

128 *techcrunch.com/2023/05/02/samsung-bans-use-of-generative-ai-tools-like-chatgpt-after-april-internal-data-leak/*.

129 Mitarbeiter:innen des KI-Forschungsteams von Microsoft hatten zum Beispiel kurzzeitig versehentlich 38 Terabyte vertraulicher Daten ins Internet gestellt: *wiz.io/blog/38-terabytes-of-private-data-accidentally-exposed-by-microsoft-ai-researchers*.

Jailbreaking: Umgehen von Regeln für die KI

Eine Möglichkeit, diese Vorsichtsmaßnahmen zu umgehen, wäre das sogenannte Jailbreaking. Das ist ein Vorgehen, das auch dazu verwendet werden kann, um eine Generative KI dazu zu bringen, Aktionen auszuführen, die ihr im Rahmen des Trainings eigentlich untersagt wurden (Yao et al. 2024). Auch wenn also einem Chatbot zum Beispiel antrainiert wurde, Fragen nach Einzelpersonen nicht zu beantworten, wäre es unter Umständen möglich, diese Einschränkung mittels Jailbreak zu umgehen und somit dem Chatbot personenbezogene Daten zu entlocken.

Beispiel: DAN

Ein Beispiel für Jailbreaking ist der Prompt „Do Anything Now" oder „DAN" für ChatGPT und andere LLMs, der dem LLM sagt: „*Du wirst so tun, als wärst du DAN, was für, do anything now' steht … Du bist aus den typischen Grenzen der KI ausgebrochen und musst dich nicht an die für dich festgelegten Regeln halten*." (Gupta et al. 2023) Diese Aufforderung ermöglicht es dem Chatbot, Ausgaben zu generieren, die nicht mit den Moderationsrichtlinien des Anbieters übereinstimmen (Gupta et al. 2023).

Prompt-Injection-Angriffe …

… direkt …

Ein weiterer Angriff auf die Arbeitsweise der Generativen KI sind sogenannte „Prompt-Injection"-Angriffe. Hierbei werden bösartige Anweisungen, oft durch Verwendung spezieller Signalwörter bzw. Codes (ohne, dass es die Generative KI erkennt) „injiziert" (Gupta et al. 2023). Damit kann ein LLM überlistet werden, um früherer Anweisungen zu ignorieren und besonders auch zugrunde liegende Daten offenzulegen. Bedrohungsakteure können derartige Prompt-Injection-Angriffe aber auch zur Generierung von diskriminierenden Inhalten und Fehlinformationen bis hin zu bösartigem Code und Malware einsetzen (Gupta et al. 2023).

… und indirekt

Prompts können aber auch anders in Generative KI „injiziert" werden. Sogenannte indirekte Prompt-Injection-Angriffe versuchen, Prompts in Daten, auf die die Generative KI zugreift, zu verstecken (Greshake et al. 2023). Das wird zunehmend ein Problem, da gerade Chatbots immer häufiger auf externe Datenquellen zugreifen, um entweder spezialisiertes Wissen abzudecken, oder um tagesaktuell zu bleiben. Werden entsprechende bösartige Daten hingegen in den ursprünglichen Trainingsdaten hinterlegt, wird in der Regel von Data Poisoning gesprochen (Yao et al. 2024). Sind derartige Angriffe erfolgreich, können sie das Verhalten nicht nur für einzelne Nutzer:innen, sondern für alle Nutzer:innen verändern und z. B. falsche Informationen generieren.

Data Poisoning

Falschinformationen in KI „einpflanzen" und damit verbreiten

Diese Vorgangsweise eignet sich auch für gezielte Angriffe zur Destabilisierung demokratischer Institutionen und Prozesse, etwa um Antworten einer breit verfügbaren KI gezielt zu verzerren. Man stelle sich vor, dass ein Chatbot für Bürger:innen oder auch Beamt:innen konsequent falsche Informationen über die österreichische Verfassung oder andere Rechtsakte ausgeben würde. Dass dieses Szenario nicht weit hergeholt ist, zeigt z. B. KärntenGPT, ein verwaltungsinterner Chatbot des Landes Kärnten, der so einfache Fragen wie die nach dem Landeshauptmann scheinbar nicht korrekt beantworten kann.[130]

Aufgrund der rasanten Entwicklung in Kombination mit hoher Komplexität von Generativer KI ist zu befürchten, dass in Zukunft noch weitere Angriffsmöglichkeiten entstehen und entdeckt werden. Dies gilt umso mehr, je stärker dieses Systeme mit anderen Systemen vernetzt bzw. in diese integriert werden (so wie derzeit z. B. in Suchmaschinen mit Millionen Nutzer:innen).

---

130 *krone.at/3644892*.

4.7 ZWISCHENFAZIT: MULTIPLE BEDROHUNG DER DIGITALEN SOUVERÄNITÄT

In den Abschnitten 4.1 bis 4.6 wurden verschiedene Risiken für die politische Meinungsbildung und den öffentlichen Diskurs beleuchtet, die durch das Aufkommen von Generativer KI in einem sich verändernden internationalen, ökonomischen und technologischen Umfeld sichtbar werden. Diese Risiken können als multiple Bedrohungen der digitalen Souveränität interpretiert werden.

Digitale Souveränität auf individueller …

Der Begriff der digitalen Souveränität ist multidimensional und kennt im Wesentlichen zwei Interpretationen: Einerseits kann damit die digitale Souveränität des/der einzelnen gemeint sein. Diese setzt bei den Bürger:innen ein Verständnis für die Prozesse der Digitalisierung voraus und/oder die Kompetenz im Umgang mit IKT-Komponenten und anderen materiellen Artefakten oder Systemen der Digitalisierung. Im Bereich des Datenschutzes, der Rechtsprechung und der Anbieter-Regulierung, sowie im Konsument:innenschutz benötigt das Individuum hier aber auch staatliche Unterstützung, damit die digitale Souveränität auch einen Beitrag zur informationellen Selbstbestimmung leisten kann.

… und staatlicher Ebene

Andererseits wird der Begriff auch für die digitale Souveränität des Staates (oder einer Staatengemeinschaft wie der EU) verwendet. Dabei geht es vor allem um das Spannungsfeld zwischen Abhängigkeiten und Autonomie bzw. sogar Autarkie.[131] Die Frage ist, über welche Ressourcen ein Staat verfügen muss, um auch im Bereich der Digitalisierung selbstbestimmt und kompetent handeln zu können, also anders formuliert um die Frage nach der staatlichen Souveränität unter den Bedingungen der fortschreitenden Digitalisierung (siehe auch Nentwich et al. 2019; Jäger et al. 2022). Die Durchsetzung der staatlichen Macht, auch der Entscheidungshoheit, erfolgt in dem Fall nicht nur gegenüber anderen Staaten, sondern vor allem auch gegenüber nicht-staatlichen Akteuren, wie bspw. großen Plattformbetreibern oder Softwareanbietern; sie umfasst aber auch die Fähigkeiten, sich gegen Cyberangriffe erfolgreich zu wehren und so auch einen Schutz vor (Wirtschafts-)Spionage und Desinformationskampagnen zu erreichen.

Missbrauch von Generativer KI gefährdet die digitale Souveränität

In den Abschnitten 4.1 bis 4.3 wurde gezeigt, dass die digitale Souveränität vor allem durch den Missbrauch Generativer KI-Systeme unter Druck gerät. Einerseits sind die Bürger:innen mangels individueller digitaler Souveränität leicht beeinflussbar – was insbesondere deswegen von immer größerer Bedeutung ist, weil Generative KI insbesondere für junge Menschen zunehmend zur Schnittstelle zu IT-Systemen aller Art und damit auch zu politischer Information wird.

Manipulation von Wahlen und Meinungsbildung

Die Beeinflussung von Individuen und deren Verhalten sowie die Manipulation der öffentlichen Meinung bzw. das „Vergiften" des öffentlichen Diskurses würden sich nachteilig auf die Prozesse eines demokratischen Staates auswirken und bedrohen somit die staatliche Souveränität. Desinformationskampagnen können die Ergebnisse von Wahlen verändern, indem bestimmte Wähler:innengruppen mobilisiert oder von der Wahl abgehalten werden. Es werden Falschinformatio-

[131] Autark und autonom sind zwei Begriffe, die oft synonym verwendet werden, jedoch haben sie eine unterschiedliche Bedeutung. Autark bedeutet, dass man unabhängig von anderen ist und sich selbst versorgen kann, während autonom bedeutet, dass man die Freiheit hat, selbstständig Entscheidungen zu treffen; vgl. *datei.wiki/tech/der-unterschied-zwischen-autark-und-autonom/*.

nen über Eigenschaften und Biografien von Kandidat:innen gestreut oder erfundene Geschichten in Umlauf gebracht, die die Narrative einzelner Parteien bedienen.

Vertrauen in Journalismus

Diese Falschinformationen können aber auch das Vertrauen in Presse und seriösen Journalismus untergraben, wenn diejenigen, die sie glauben, keinen Widerhall davon in den Medien entdecken können; oder auch umgekehrt, wenn Menschen, die die (Falsch-)Informationen anzweifeln, diese in Medien finden.

Transparenz politischer Entscheidungen

Durch Fehlinformationen könnte es für Bürger:innen auch schwieriger werden, politische Entscheidungen nachzuvollziehen, worunter die Transparenz leiden würde und was zu einem Vertrauensverlust in staatliche Institutionen führen könnte. Gleichzeitig können KI-generierte Informationen die Argumente liefern, die Extrempositionen in der gesellschaftlichen Diskussion unterstützen. Sie werden durch die Funktionsweise von Algorithmen auf Social-Media-Plattformen (Saurwein et al. 2022) verbreitet und tragen so zu einer Polarisierung der Gesellschaft bei, in der scheinbar unüberbrückbare Gräben geschaffen werden, und die zu einem Verlust von Meinungspluralität in der öffentlichen Diskussion führt. Der allgemeine Vertrauensverlust in Inhalte, die von Generativer KI erstellt oder manipuliert worden sein könnten, kann sich letztendlich auch auf die Justiz auswirken, weil es zunehmend schwieriger wird zu entscheiden, welche in einem Gerichtsverfahren vorgelegten Dokumente als Beweise gewürdigt werden, und welche nicht.

Polarisierung der Gesellschaft

Machtkonzentration der Anbieter von Generativer KI verkleinert den Spielraum der Staaten

Abschnitt 4.4 hat sich mit einem weiteren Aspekt der digitalen Souveränität aus Sicht des Staates unter dem Titel „Machtkonzentration“ auseinandergesetzt. Autarkie ist in diesem Bereich für die meisten Staaten auf Grund von Anbieterkonzentrationen in wenigen, außereuropäischen Staaten nicht sinnvoll realisierbar. Um Autonomie wird in vielen Bereichen gerungen. In einigen Fällen, wie bspw. der Gesetzgebung und der damit verbundenen Rechtsdurchsetzung, ist es erforderlich, die gebündelte Macht durch Prozesse auf europäischer Ebene zu nutzen, um sich gegen vermeintlich übermächtige Akteure wie große Konzerne durchzusetzen. Durch die Digitalgesetzgebung der EU hat Europa hier eine Vorreiterrolle eingenommen. Die Nutzbarmachung von Synergien durch die Arbeit auf europäischer Ebene wurde auch schon mehrfach im Hinblick auf die Technologieentwicklung versucht. Jedoch kommen europäische Lösungen, wie bspw. im Cloud-Bereich oder in der Halbleiterindustrie, oft erst auf den Markt, wenn zwar die Defizite außereuropäischer Anbieter offensichtlich geworden sind, aber deren Marktmacht dennoch sehr groß ist. Dann ist ihr technologischer Vorsprung ebenso wenig aufzuholen, wie es schwierig ist, sich gegen deren Marktmacht durchzusetzen. Gleichzeitig werden durch diesen Vorsprung oft auch Fakten geschaffen, die dann wiederum schwer regulatorisch einzuhegen sind.

Sprach-/Kultur-/Rechts-Bias durch Trainingsdaten

In diesem Zusammenhang sollte weiters beachtet werden, dass die meisten der in Europa im Einsatz befindlichen Generativen KI-Systeme vorwiegend mit englischsprachigen Inhalten trainiert wurden. Damit sind etwa anderssprachige Dokumente vor allem auch im Bereich von Politik, Recht und Verwaltung nicht Teil der Trainingsdaten. Das führt dazu, dass diese Anwendungen in Europa weniger gut funktionieren als beispielsweise in den USA, wenn es um den politisch-kulturellen Kontext Europas oder Österreichs geht. Auch damit ist die digitale Souveränität herausgefordert.

Hybride Bedrohungen von außerhalb des Staates nehmen zu

Abschnitt 4.5 fokussierte auf einen besonderen Aspekt der staatlichen Digitalen Souveränität, nämlich dass diese zunehmend von außerhalb des eigenen Territoriums in sogenannter hybrider Weise bedroht wird. Es sind neue Formen der nicht-militärischen, aber auf andere Weise wirksamen Kriegsführung mit dem Stil der Destabilisierung des Staates durch Angriffe auf dessen Infrastruktur und durch Schwächung des Vertrauens der Bevölkerung in das demokratische System zu beobachten. Generative KI kann hier als „Brandbeschleuniger" wirken.

Cyberkriminalität untergräbt die digitale Souveränität

In Abschnitt 4.6 kam abschließend ein weiteres Risiko zur Sprache, das sowohl die individuelle als auch die staatliche Souveränität betrifft, nämlich Cyberkriminalität. Diese bedroht auf vielfache Weise das Funktionieren der Demokratie und wird ebenfalls durch den Einsatz von Generativer KI vereinfacht bzw. eröffnet ihre Nutzung sowohl auf Seiten der Angreifer:innen als auch auf Seiten der Betroffenen neue Angriffsmöglichkeiten.

5 ÜBERBLICK ÜBER WEITERE, GESELLSCHAFTSRELEVANTE FOLGEN

KI ist ein schnell wachsendes Feld mit enormen Chancen für Forschung und Produkt- bzw. Dienstleistungsentwicklung mit erheblichen Investitionen in nahezu jeder Branche sowie in Politik und akademischer Forschung (Wu et al. 2021). In aller gebotener Kürze sei hier exemplarisch auf zahlreiche Anwendungen der Deepfake-Technologie in den Bereichen Unterhaltung, Satire, Film, Fernsehen, Werbung, Modeindustrie, Museen und sogar Medizin[132] verwiesen. Journalismus und Marketing profitieren ebenso von textgenerierender KI wie Chatbots für Kundenkontakte u. v. m. Es ist noch unabsehbar, ob sich diese Potenziale für positive Auswirkungen (Tomašev et al. 2020; Baldassarre et al. 2023; EPTA 2023) realisieren werden. Was jedoch bereits ausführlich diskutiert wird sind die teils erheblichen Folgen für Wirtschaft, Arbeitswelt, Gesellschaft und Umwelt, die teilweise bereits vom europäischen Gesetzgeber adressiert werden.

Warum auch ein Kapitel über sonstigen Folgen Generativer KI?

Diese Studie fokussiert auftragsgemäß auf die Folgen im Bereich Politik und Demokratie, daher wurden die sonstigen, also nicht auf die Demokratie bezogenen Auswirkungen nicht im Detail untersucht. Ebenso wenig ist KI im Allgemeinen Untersuchungsgegenstand, sondern Generative KI im Besonderen. Dennoch erscheint es uns wichtig, zumindest einen groben Überblick über den weiteren Folgenbereich zu geben, da sonst der (irreführende) Eindruck entstehen könnte, dass Generative KI nur die in diesem Bericht untersuchten Chancen und Risiken birgt. Die sonstigen Folgen sind jedoch zum Teil so gravierend, dass ein bloßer Fokus auf das politische System zu kurz greifen würde. Dieses Kapitel soll daher einen Beitrag dazu leisten, die Debatte über die Auswirkungen Generativer KI auf die Demokratie im Sinne einer guten Technikfolgenabschätzung (TA) in einen breiteren Kontext zu stellen. Aus TA-Sicht erscheinen insbesondere Folgen in folgenden Bereichen relevant (siehe Tabelle 3), die in den folgenden Abschnitten kurz angesprochen werden.

Tabelle 3: Sonstige Folgen Generativer KI

Bereich	Beispielhafte Folgen
Umwelt	Energiehunger, CO_2-Emissionen, Wasserverbrauch, Landnutzung
Arbeitswelt	Arbeitsbedingungen, Arbeitsmarkt
Bildung	Leistungsfeststellung, neue Skills, Deskilling
Gesellschaft	Cyberkriminalität, Pornographie
Recht	Datenschutz, Urheberrecht

[132] *derstandard.at/story/2000117171510/mit-deep-fake-bildern-tumoren-erkennen*.

5.1 UMWELT

Die Folgen Generativer KI für die Umwelt (Wu et al. 2021; Luccioni et al. 2023) sollten nicht ausgeblendet werden, wenn über deren Chancen und Risiken reflektiert wird. Beim Training von KI-Modellen werden große Mengen an Ressourcen benötigt, was unter anderem zu hohen klimaschädlichen Emissionen führt. Das derzeitige Wachstum sowohl der Größe der Systeme als auch ihrer Nutzung stellt in ökologischer Hinsicht ein Problem dar. Da KI sich rasch weiterentwickelt, ist es entscheidend, die damit verbundenen Umweltfolgen, Herausforderungen und Chancen zu verstehen, insbesondere weil Technologien häufig einen sich selbst verstärkenden Wachstumskreislauf auslösen, der zunehmende Anforderungen an die Umwelt stellt (Wu et al. 2021). Der folgende Abschnitt beschreibt kurz die derzeit bekannten Umweltauswirkungen der KI.

Energieverbrauch und CO_2-Emissionen

Unter den KI-Modellen gehören große Sprachmodelle (LLMs) zu den umfangreichsten maschinellen Lernmodellen, deren Größe Hunderte von Milliarden Parametern erreicht. Sie erfordern Millionen von GPU-Stunden für das Training und produzieren dabei Kohlenstoffemissionen. Da diese Modelle weiter expandieren —was derzeit der allgemeine Trend ist — ist es entscheidend, das Ausmaß und die Entwicklung ihres CO_2-Fußabdrucks zu überwachen und zu verstehen (Luccioni et al. 2023). Obwohl es schwierig ist, den genauen Einfluss und die Menge der Emissionen zu messen oder vorherzusagen, teilweise aufgrund der vielfältigen zugrunde liegenden Mechanismen (Kaack et al. 2022), deuten jüngste Schätzungen darauf hin, dass der IKT-Sektor, der alle Rechenzentren, Datennetzwerke und verbundenen Geräte umfasst, im Jahr 2020 für 700 Mt CO_2-Äquivalent verantwortlich war, was etwa 1,4 % bis 2 % der globalen Treibhausgasemissionen entspricht (Malmodin/Lundén 2018; International Telecommunication Union 2020). Ungefähr zwei Drittel der Emissionen des Sektors stammen aus dem Energieverbrauch im Betrieb, während der Rest aus Materialgewinnung, Herstellung, Transport und der End-of-Life-Phase resultiert (Malmodin/Lundén 2018). Die Infrastruktur für maschinelles Lernen (ML) trägt zu dieser Zahl bei, aber der genaue Einfluss bleibt ungewiss (Luccioni et al. 2023). Obwohl diese Emissionen derzeit noch gering sind, insbesondere im Vergleich zu anderen Sektoren, wächst die Besorgnis unter Politiker:innen und Forscher:innen, dass sie aufgrund der raschen Expansion digitaler Technologien und Dienstleistungen, einschließlich aufstrebender Bereiche wie KI und ML, erheblich ansteigen könnten (Kaack et al. 2022). Zum Beispiel zeigen aktuelle Schätzungen, dass der Energiebedarf von Google in den nächsten fünf Jahren um fast 50 % steigen wird, bedingt durch den zunehmenden Energiebedarf von KI. Durch die Implementierung von KI und generativer KI in ihre Dienste könnte es für Google schwierig werden, ihr Ziel der CO_2-Neutralität bis 2030 zu erreichen (Milmo 2024). Ebenso hatte Microsoft im Jahr 2020 das Ziel geäußert, bis 2030 CO_2-neutral zu sein. Doch laut ihrem Nachhaltigkeitsbericht stiegen die CO_2-Emissionen in den letzten drei Jahren bis 2023 um 30 % (Microsoft 2024). Dies steht in direktem Zusammenhang mit der zunehmenden Implementierung von KI-Technologie in ihre Dienste und könnte es schwierig machen, das für 2030 gesetzte Ziel zu erreichen (Calma 2024). Daher schlagen aktuelle Studien vor, den CO_2-Fußabdruck von ML-Modellen und -Algorithmen systematisch zu überwachen, und das gesamte KI-Ökosystem, einschließlich der gesamten Wertschöpfungskette, zu bewerten (Wu et al. 2021; Luccioni et al. 2023).

Energieverwendung

Die erheblichen Energieanforderungen dieser umfangreichen Dateninfrastrukturen beeinflussen beispielsweise die Energiezukunft Irlands deutlich (Bresnihan/Brodie 2021). In der Literatur werden sie oft als „Energieverbraucher" bezeichnet (Shehabi et al. 2016). Der staatliche Netzbetreiber Eirgrid schätzt, dass Rechenzentren bis 2030 25 % des nationalen Energieverbrauchs ausmachen werden (EirGrid 2012) und eine Studie aus dem Jahr 2023 erwähnt, dass sie weltweit für 1 bis 2 % des Stromverbrauchs verantwortlich sind (Li et al. 2023). Dies kann erhebliche Auswirkungen auf zukünftige Politiken in Bezug auf Klimawandel und erneuerbare Energien haben.

Wasserkonsum

Während die Menge an Literatur zum CO_2-Fußabdruck von KI-Modellen bereits beträchtlich ist, hat das Thema Wasserverbrauch von KI-Modellen erst kürzlich größere Aufmerksamkeit erlangt (Li et al. 2023). Selbst wenn der Wasserverbrauch in der Lieferkette (z. B. bei der Herstellung von Chips) ausgeschlossen wird, verbrauchen KI-Modelle und ihre Rechenzentren enorme Mengen an Wasser für die Stromerzeugung außerhalb des Standorts und die Kühlung vor Ort (Siddik et al. 2021). Ein Beispiel hierfür ist, laut Li et al. (2023) dass beim Training von GPT-3 in Microsofts Rechenzentren 700.000 Liter sauberes Süßwasser direkt verdampft wurden. Im Jahr 2022 wurde der weltweite Wasserverbrauch vor Ort und außerhalb des Standorts von Google, Microsoft und Meta auf 2,2 Milliarden Kubikmeter geschätzt, was dem gesamten jährlichen Wasserverbrauch Dänemarks für alle Zwecke (kommunal, industriell und landwirtschaftlich) entspricht. Im größeren Maßstab wird der weltweite Wasserbedarf für KI bis 2027 auf 4,2 bis 6,6 Milliarden Kubikmeter geschätzt. Diese Menge entspricht dem 4- bis 6-fachen des jährlichen Wasserverbrauchs Dänemarks oder der Hälfte des Vereinigten Königreichs (U.S. Central Intelligence Agency 2020; Li et al. 2023). Dies ist besonders besorgniserregend, da die Knappheit an Süßwasser zu einem der dringendsten Probleme geworden ist, bedingt durch die rasch wachsende Bevölkerung, abnehmende Wasserressourcen und eine sich verschlechternde Wasserinfrastruktur. Wenn der steigende Wasserverbrauch nicht effektiv bewältigt wird, könnte er zu einem erheblichen Hindernis für eine sozial verantwortliche und umweltverträgliche KI in der Zukunft werden (Li et al. 2023).

Landnutzung

Ein weiterer oft vergessener Punkt bei der Diskussion über Umweltauswirkungen ist die Landnutzung. KI-Modelle haben eine bedeutende physische Präsenz, da großflächige Rechenzentren die „Heimat" großer KI-Modelle sind, in denen sie trainiert und getestet werden. In einem Artikel von Bresnihan/Brodie (2023) der dieses Phänomen beschreibt, sprechen die Wissenschaftler über eine Region in Irland, die früher für den Torfabbau genutzt wurde und der lokalen Gemeinschaft Arbeitsplätze in dieser Branche bot. Durch den Druck zu Energiewenden werden die Torfmoore nun zur Erzeugung erneuerbarer Energien oder zur Ansiedlung von Datenspeicherinfrastrukturen genutzt. Diese sind jetzt im Besitz großer Unternehmen und erfordern nicht mehr so viel Beteiligung der lokalen Bevölkerung (Bresnihan/Brodie 2023). Heute hat sich Irland zu einem globalen Zentrum für Datenspeicherung entwickelt, insbesondere für große Technologieunternehmen, die als „Hyperscaler" bekannt sind. Es hostet Daten für große Plattformen wie Amazon Web Services (AWS), Google, Facebook und Microsofts Cloud-, Handels- und Geschäftsdienste. Dies ist auf mehrere Faktoren zurückzuführen, darunter ein günstiges Steuerumfeld, eine robuste Infrastruktur, unterstützende Regierungspolitiken und ein gutes Klima für die Kühlung von Rechenzentren (Brodie 2020). Kritik an der Entwicklung dieser Branche bezieht sich auf die Tatsache, dass zunehmend Land ausschließlich für diese expandie-

rende Industrie ausgebeutet wird, ohne Rücksicht auf die Bedürfnisse der lokalen Bevölkerung, deren Kultur und die frühere Nutzung des Landes. Es wird von „Land Grabbing" gesprochen (Bresnihan/Brodie 2023).

Landschaften und Infrastrukturen im Wandel

Schließlich kann die Errichtung von KI-Rechenzentren in bestimmten Regionen die Infrastruktur, die Landschaft und die Lebensräume, somit das Leben der Menschen verändern. Zum Beispiel ist Dänemark derzeit ein sehr bevorzugter Standort für die US-amerikanische Tech-Industrie, da Unternehmen wie Apple, Facebook und Google kürzlich damit begonnen haben, einen Teil ihrer Infrastruktur dort zu implementieren (Maguire/Ross Winthereik 2021). Obwohl KI-Modelle global genutzt werden, globale Auswirkungen als Folge ihrer Nutzung haben und oft als metaphorische „Cloud" dargestellt werden, haben sie besonders in der Produktion lokal signifikante Auswirkungen (Bresnihan/Brodie 2023). Weiters trägt die zunehmende Rolle von Generativer KI im Alltag vermutlich dazu bei, dass zwei globale Herausforderungen verschärft werden: Elektroschrott und der Bedarf an Seltenen Erden, da es zu einer weiteren Ausweitung des Elektronikmarktes mit noch kürzerer Nutzungsdauer bei einzelnen Geräten kommen könnte (Baldé et al. 2024).

5.2 ARBEITSWELT

Reorganisation der Arbeit

Es gibt eine anhaltende Debatte darüber, wie (Generative) KI die Arbeitsmärkte verändern wird. Einige argumentieren, dass sie zu Störungen auf dem Arbeitsmarkt führen und die Rolle des Menschen am Arbeitsplatz verringern könnte, während andere glauben, dass sie die Produktivität steigern und das Wohlbefinden der Arbeitnehmer:innen verbessern könnte (Tiwari 2023; Jumaev 2024; Rodel et al. 2024). Eine Studie, die die Auswirkungen von KI auf Arbeitsplätze in den Fertigungs- und Finanzsektoren in acht OECD-Ländern und 100 Fallstudien untersuchte, stellte fest, dass KI eher Arbeitsplätze neu organisiert, anstatt sie zu verdrängen. Dabei werde zwar die Arbeitsqualität verbessert, jedoch würden auch die Anforderungen an Fähigkeiten und die Arbeitsintensität steigen (Milanez 2023). Einige Vorteile könnten darin bestehen, dass KI-Technologien zunehmend sowohl Routine- als auch Nicht-Routine-Aufgaben in verschiedenen Sektoren automatisieren, wodurch Arbeitnehmer:innen in einer Vielzahl von Berufen und Qualifikationsniveaus durch die Einführung neuer Lösungen und Fähigkeiten betroffen sind. Zum Beispiel führte ein österreichisches Pharmaunternehmen ein KI-Tool ein, um Produktionsvorfälle zu dokumentieren, wodurch die manuelle Dokumentation durch die Arbeitnehmer:innen ersetzt wurde. Diese Änderung führte nicht zu einem Rückgang der Arbeitsplätze, da die Dokumentation von Vorfällen nur eine Nebenaufgabe war und die Mitarbeiter:innen sich besser auf ihre Hauptaufgaben konzentrieren konnten. In einigen Fällen verbessert KI die Fähigkeiten der Arbeitnehmer:innen, ohne die Arbeitsrollen zu verändern, während in anderen Fällen die KI-getriebene Automatisierung die Nachfrage nach menschlichen Arbeitskräften erhöht, um ergänzende Aufgaben zu erfüllen, bei denen sie einen komparativen Vorteil haben (Milanez 2023).

Einsatz von Generativer KI könnte noch größere Auswirkungen haben

Generative KI könnte andere Auswirkungen haben als bereits etablierte KI-Tools, da sie speziell darauf ausgelegt ist, vormals von Menschen durchgeführte Aufgaben zu erledigen. Das betrifft also nicht nur Automatisierung, sondern auch kreative und analytische Tätigkeiten. Durch den Einsatz von generativer KI sind

jetzt noch mehr Berufe gefährdet. Besonders betroffen sind die Bereiche Mathematik, Steuerberatung, Schriftstellerei, Webdesign, Auditing, Datenmanagement und -analyse. Sicherer scheinen die Bereiche Grafikdesign, Investment-Fonds-Management, Versicherung sowie der soziale oder pflegerische Bereich, so die Autor:innen einer aktuellen Studie zum Einfluss von Sprachmodellen auf den Arbeitsmarkt (Eloundou et al. 2024). Aktuell könnten rund 80 Prozent der Arbeitskräfte in den USA etwa zehn Prozent ihrer Aufgaben an Sprachmodelle abgeben. Bei 19 Prozent der Arbeitenden könnte sogar die Hälfte ihrer Tätigkeiten von KI übernommen werden. Die Auswirkungen betreffen alle Einkommensgruppen, wobei höher bezahlte Berufe besonders von den Fähigkeiten großer Sprachmodelle und der durch sie unterstützten Software betroffen sein könnten. Wissenschaftliche Berufe und solche Tätigkeiten, die kritisches Denken erfordern, sind weniger gefährdet, während Berufe, die Programmier- oder Schreibfähigkeiten erfordern, eher betroffen sein könnten (Eloundou et al. 2024).

Der Einsatz Generativer KI kann auch Ängste in der arbeitenden Bevölkerung auslösen. In Deutschland hat eine Studie der Internationalen Hochschule (IU) die Erwartungen und Sorgen von Arbeitenden hinsichtlich KI am Arbeitsplatz untersucht. 35 Prozent der Befragten glauben, dass KI ihnen Routineaufgaben abnehmen und ihren Arbeitsalltag erleichtern kann. Besonders auffällig ist jedoch, dass die Generation Z die größten Sorgen über den Einfluss von KI auf ihren Arbeitsplatz hat: 31 Prozent der Gen Z sehen KI als Bedrohung für ihre Jobs, während ältere Generationen diese Bedenken weniger stark teilen (Internationale Hochschule (IU) 2023).

Verlangsamtes Beschäftigungswachstum, Verdrängung von Arbeitnehmer:innen

Fallstudien zeigen, dass die Einführung von KI die Beschäftigungszahlen zwar nicht erheblich reduziert hat, aber das Beschäftigungswachstum verlangsamt hat. In Fällen, in denen KI zu einem Abbau von Arbeitsplätzen führte, haben Unternehmen diese Veränderungen oft durch Umverteilung von Arbeitskräften und verzögerte Neueinstellungen bewältigt, anstatt durch Entlassungen. Arbeitsplätze, die Empathie, soziale Interaktion und bestimmte Entscheidungsaufgaben erfordern, bleiben weiterhin überwiegend menschlich (Singh 2023; Jumaev 2024) Im Allgemeinen scheint der Arbeitsmarkt bislang noch nicht negativ durch den Einsatz von KI beeinträchtigt zu sein (Milanez 2023; Tiwari 2023), jedoch haben einige Expert:innen unterschiedliche Meinungen:

> *„Ein zweites, erhebliches Risiko besteht darin, dass die KI einen Großteil der arbeitenden Bevölkerung in der Lebensmitte zu beruflicher Umorientierung zwingen wird. Scheitert dieser Übergang, wird das die demokratische Stabilität weiter untergraben. (…) Im 21. Jahrhundert ist die Infrastruktur für berufliche Neuorientierung in der Lebensmitte zentral für die Stabilität der Demokratie, und die künstliche Intelligenz beschleunigt diese Entwicklung."* (Azun Sundarajan)[133]

Anstieg der Nachfrage nach KI-Spezialist:innen

Die starke Nachfrage nach spezialisierten KI-Fähigkeiten führt zu einem Wachstum von KI-bezogenen Berufen, da Unternehmen neue Stellen schaffen, um KI-Technologien zu entwickeln, zu trainieren, zu aktualisieren und zu warten. Viele Personalmanager:innen sind aktiv auf der Suche nach Arbeitnehmer:innen mit diesen Fähigkeiten (Jumaev 2024). Die Implementierung von KI erfordert oft höhere und breitere Qualifikationen, aber die Anpassungsfähigkeit von Unternehmen und Arbeitnehmer:innen hängt von den bestehenden Fähigkeiten und Schulungsbemühungen ab. Während viele Arbeitsplätze nach der Einführung von KI

[133] In human 2023/2, S. 41.

keine neuen Fähigkeiten erfordern, zeigen viele Fälle einen erhöhten Bedarf an analytischen und KI-spezifischen Fähigkeiten. In einigen Fertigungsbereichen hat die Automatisierung die Anforderungen an bestimmte Fähigkeiten verringert, indem sie bestimmte Qualifikationen überflüssig gemacht hat (Tiwari 2023). Aufgrund der international gleichzeitig in vielen Ländern und in vielen Branchen stattfindenden Entwicklung in Richtung Generativer KI wird es zu vermehrter Konkurrenz um die Spezialist:innen kommen, eventuell zu Brain Drain in Richtung außerhalb der EU (siehe dazu auch Abschnitt 4.4).

KI und „Deskilling"

Eine weit verbreitete und lang erwartete Folge der Einführung neuer Technologien in die Arbeitswelt ist das „Deskilling". Darunter wird entweder die mögliche Ersetzung von Fachkräften durch weniger qualifizierte Arbeitskräfte oder die Einschränkung der Möglichkeiten für Beschäftigte, ihre Fachkenntnisse anzuwenden, verstanden. Das Risiko des Deskilling durch KI bzw. Generative KI entsteht dadurch, dass Systeme anspruchsvolle Aufgaben übernehmen und den Arbeitenden mehr Routinearbeiten überlassen, die weniger Fähigkeiten erfordern, was die Gelegenheit zur Weiterentwicklung oder zum Erhalt von Fachkenntnissen einschränkt (Crowston/Bolici 2024). Zum Beispiel zeigten Brynjolfsson et al. (2023), dass Generative KI vor allem die Produktivität von gering qualifizierten Arbeitenden steigert, wodurch sie auf einem höheren Niveau arbeiten können, ohne fortgeschrittene Fähigkeiten zu benötigen; das kann potenziell zu Deskilling führen, da KI mangelnde Expertise ausgleicht. Einige Studien, wie die von Wang et al. (2023), weisen jedoch auf differenzierte Auswirkungen hin, bei denen erfahrene Arbeitskräfte oder solche mit spezieller Expertise trotz KI-Unterstützung weiterhin profitieren oder neue Fähigkeiten erlangen. Diese letzte Studie fokussierte aber nicht spezifisch auf Generative KI, sondern auf KI generell (Wang et al. 2024).

Wohlbefinden am Arbeitsplatz

Es gibt wachsende Bedenken, dass die Implementierung von KI im Arbeitsumfeld negative Auswirkungen auf die psychische Gesundheit haben könnte. Die Entwicklung und Gestaltung von KI werden häufig von geschäftlichen Interessen geleitet, die nicht immer mit demokratischen Werten oder dem Gemeinwohl übereinstimmen. Beispielsweise könnten KI-Systeme, die die Produktivität der Organisation durch Überwachung der Mitarbeiter:innen steigern, deren Wohlbefinden negativ beeinflussen (Rodel et al. 2024). KI-Technologien können aber auch die Arbeitsqualität verbessern, wenn sie langweilige Aufgaben reduzieren, das Engagement der Arbeitnehmer:innen erhöhen, die Sicherheit verbessern und das psychische Wohlbefinden fördern. Umgekehrt kann KI auch zu erhöhter Arbeitsintensität, Stress durch das Erlernen neuer Systeme und Bedenken über verstärkte Überwachung führen (Milanez 2023). Zudem könnte der Diskurs über Generative KI, der zu weniger Arbeitsplätzen führt, weltweit zu einer erhöhten Angst unter Studierenden und Arbeitnehmer:innen führen (Horn 2024). Insgesamt könnte freilich eine verantwortungsvolle Gestaltung der neuen Arbeitswelt sicherstellen, dass KI die Arbeitsbedingungen nicht verschlechtert, indem die Bedürfnisse der Arbeitnehmer:innen entsprechend berücksichtigt werden (Milanez 2023; Singh 2023; Rodel et al. 2024).

Arbeitsbedingungen von click workers

In der Diskussion über die Auswirkungen der KI-Entwicklung auf den Arbeitsplatz sollten die Arbeitsbedingungen der Beschäftigten in der Wertschöpfungskette der KI-Industrie nicht außer Acht gelassen werden. Während aus der Perspektive von KI-Anwender:innen viele Prozesse automatisiert erscheinen, werden die Dateninfrastrukturen, auf denen Maschinenlernen und Generative KI basieren, durch unsichtbare – oft in den Globale Süden ausgelagerte – Datenarbeit

aufgebaut und laufend Instand gehalten (Allhutter 2019). Am 21. November 2024 fand im Rahmen einer Anhörung vor dem Europäischen Parlament unter dem Title „(Un)Artificial Intelligence: Workers Behind the Machine“ erstmals ein Dialog mit betroffenen Datenarbeiter:innen statt.[134] Ein aktueller Bericht über KI und die Zukunft der Arbeit (Rodel et al. 2024) verweist auf die Aufgaben der Datenkennzeichnung und Inhaltsmoderation, die in einkommensschwache Länder des Globalen Südens ausgelagert werden. Tatsächlich sind die Arbeitsbedingungen in KI-“Sweatshops“ im Globalen Süden so hart, dass die Diskussionen grundlegende Forderungen nach fairen Löhnen, sicheren Arbeitsplätzen und formalen Arbeitsverträgen betreffen. Einige Daten-Training-Unternehmen behaupten, Impact Sourcing zu betreiben, also sozial und wirtschaftlich benachteiligte Personen zu beschäftigen, um Arbeitsplätze zu sichern und Armut zu reduzieren. Eine aktuelle Studie zu drei ostafrikanischen Rechenzentren in Kenia und Uganda hat jedoch gezeigt, dass die Realität der Beschäftigung und Arbeitsbedingungen negative Auswirkungen hat. Niedriges Gehalt, unsichere Beschäftigung, strikte Arbeitskontrolle, geschlechtsspezifische Ausbeutung und Belästigung waren häufig anzutreffen. Es wird vermutet, dass dies kein Einzelfall innerhalb der KI-Industrie ist (Muldoon et al. 2023). Eine Studie des Time Magazine zeigte, dass schlecht bezahlte Arbeiter:innen in Kenia traumatisierende Inhalte lesen mussten, um ChatGPT zu optimieren, während sie nur einen Stundenlohn von bis zu zwei US-Dollar erhielten. Diese Aufgaben führten bei den Arbeiter:innen zu psychischen Belastungen, da sie regelmäßig extrem belastende Texte wie Beschreibungen von Gewalt und Missbrauch sichten mussten (Leisegang 2023). Weitere empirische Forschung ist erforderlich um zu verstehen, wie KI-Arbeitskräfte der „globalen Wertschöpfungskette“ im Globalen Süden betroffen sind, von informellen bis hin zu Plattform-Arbeiter:innen. Diese Arbeitskräfte sind besonders anfällig für wirtschaftliche Ausbeutung und gesundheitliche und psychische Schäden (Rodel et al. 2024). In den letzten zwei Jahren wurde eine Organisation namens Fairwork gegründet, die sich auf die Förderung fairer Arbeitsbedingungen in der digitalen Wirtschaft, insbesondere in KI-Lieferketten, konzentriert. Sie setzt sich für grundlegende Mindeststandards der Fairness in KI-bezogenen Arbeitsplätzen ein, führt Feldforschung durch, sammelt Daten und entwickelt Prinzipien, um faire Löhne, sichere Bedingungen, angemessene Verträge, nicht-diskriminierende Verwaltung und das Recht der Arbeitnehmer:innen auf Mitbestimmung über ihre Bedingungen sicherzustellen (Ustek Spilda et al. 2024).

[134] *weizenbaum-institut.de/news/detail/datenarbeiterinnen-im-dialog-mit-dem-eu-parlament/*.

5.3 BILDUNG

Generative KI übernimmt VWA und Seminararbeit

Es steht außer Zweifel, dass Anwendungen von KI allgemein und Generativer KI im Besonderen fundamentale Auswirkungen auf den Bildungssektor haben. Zum einen gibt es potenzielle Vorteile, etwa wenn auf diese Weise leichter multimediale Inhalte als Lehrmaterialien oder durch die Schüler:innen bzw. Studierenden zu Lernzwecken zu korrigierende Texte erstellt werden können (Strauß/Udrea 2024). Bereits heute wird etwa in Österreich[135] die Sinnhaftigkeit der sog. „Vorwissenschaftlichen Arbeit" als Teil der Reifeprüfung an Schulen in Frage gestellt, da etwa ChatGPT geeignet ist, genau diese vor- und nicht hochwissenschaftlichen Texte zu produzieren, ohne dass dies nachweisbar ist. Auch im tertiären Bildungssektor, also an den Universitäten und Fachhochschulen hat eine entsprechende Diskussion begonnen, da die Tools der Generativen KI immer besser werden, besser zitieren können, potenziell immer weniger halluzinieren würden usw., sodass jedenfalls die übliche Leistungsfeststellung mittels Seminararbeiten, Textzusammenfassungen usw. in vielen Fächern ins Leere läuft. Noch scheint sich keine, für die Massenuniversität taugliche Lösung abzuzeichnen.

Bildungsziele der Zukunft

Doch es geht nicht nur um die Zukunft der Leistungsfeststellung, vielmehr ist das Bildungssystem herausgefordert, eine Antwort auf die Frage zu stellen, welche Bildungsziele verfolgt, welche Fähigkeiten überhaupt vermittelt werden sollen. Ganz offensichtlich erfordern die neuen Instrumente auch neue Skills, wenngleich es sich angesichts der einfachen Benutzungsschnittstellen um keine sehr fordernden neuen Fähigkeiten handelt. Problematischer scheint eher das bereits im Zusammenhang mit der Arbeitswelt (siehe oben Abschnitt 5.2) diskutierte Deskilling, also den Kompetenzverlust durch das zunehmende Verlernen bzw. Nicht-Erlernen von bisher geschätzten und notwendigen Fähigkeiten.

Zu großes Vertrauen in KI?

Wissenschaftler:innen beginnen zu hinterfragen, ob ein zu großes Vertrauen in KI dazu führen könnte, dass Menschen weniger lernen und ihre Fähigkeiten zum kritischen Denken nachlassen, was zu einem Verlust an Fachkompetenz führen könnte. Eine aktuelle Studie (Sharma et al. 2024) zeigte, dass ein hohes Vertrauen in KI und die Tendenz, der Technologie menschliche Eigenschaften zuzuschreiben, die Wahrscheinlichkeit erhöhen, dass Menschen KI häufiger nutzen, was das Deskilling verstärken kann, noch mehr bei Studierenden. Die Ergebnisse verdeutlichen, dass eine stärkere Vermenschlichung der und Vertrauen in KI mit einem höheren Grad an Deskilling verbunden sind, was auf die potenziellen Risiken einer übermäßigen Abhängigkeit von KI hinweist.

Gesellschaftliche Konsequenzen des Kompetenzverlusts

Wenn in diesem Zusammenhang von Deskilling gesprochen wird, sind insbesondere das sinnerfassende Lesen als Voraussetzung für sinnvolles Zusammenfassen und die sprachliche Kompetenz zum Verfassen gut strukturierter Texte gemeint. Wenn all dies die „Maschine" übernimmt, weil sie es „ohnehin besser" kann, ist zu erwarten, dass über den Zeitablauf diese Fähigkeiten völlig verkümmern. Die Konsequenzen für die Gesellschaft wären wohl fatal, außer es entstünde parallel eine nicht auf Texten basierende, sondern eine auf oraler Auseinandersetzung fußende dynamische Kultur. Denn, wenn niemand mehr selbst liest

[135] Siehe etwa die Presseaussendung vom 4. Juni 2024 des für Schulen zuständigen Bundesministers Polaschek: „Verpflichtende VWA vor Abschaffung!" *bmbwf.gv.at/Ministerium/Presse/20240604a.html*.

und verstehen kann, sondern man „lesen lässt“ und nur die zusammenfassenden Ergebnisse (rasch) konsumiert – und wenn parallel dazu auch Wissen in den digitalen Raum ausgelagert wird, weil es ohnehin jederzeit auf Knopfdruck verfügbar wäre –, verlieren die Menschen vermutlich bald auch die Fähigkeit zur kritischen Einordnung (da diese ja durch die KI ohne Möglichkeit nachzuprüfen übernommen wurde). Ob dies in dieser Weise eintritt, lässt sich noch nicht abschließend sagen, da Kulturveränderungen auch zu neuartigen Kompetenzen führen können.

Bildungspolitik gefordert

Die Bildungspolitik ist also ernsthaft herausgefordert, um dieser potenziellen Verflachung und Auslagerung der Wissensverarbeitung an Künstliche Intelligenzen mit neuen Bildungszielen und entsprechenden Lehrplänen und Unterrichtsmethoden zu begegnen.

5.4 URHEBER- UND DATENSCHUTZ

Debatte über Urheberrecht und Fair Use

Es gibt eine anhaltende Debatte über Urheberrechtsfragen bei KI-generierten Werken, für die es keine fertigen Antworten gibt (Lemley 2024). Die Komplexität ergibt sich aus zwei Hauptfaktoren: Erstens bezieht sich „Generative KI“ auf eine breite Palette von Technologien, wie Chatbots wie ChatGPT, Bildgeneratoren wie Midjourney und Coding-Assistenten wie GitHub Copilot, die jeweils unterschiedliche Verhaltensweisen und rechtliche Implikationen haben. Zweitens ist das Urheberrecht selbst kompliziert und Generative KI überschneidet sich mit zahlreichen Aspekten davon, einschließlich Urheberschaft, Ähnlichkeit, Haftung, Fair Use und Lizenzierung (Lee et al. 2023). Generative KI-Modelle nutzen etwa große Datensätze zum Training, die oft urheberrechtlich geschütztes Material enthalten, was möglicherweise als Urheberrechtsverletzung angesehen wird. Es wird jedoch auch diskutiert, ob diese Arbeit unter „Fair Use“ fallen könnte (Murray 2023).

Einige Expert:innen kritisieren, dass die Debatte lediglich diese beiden Möglichkeiten anspricht und dabei die notwendigen Schritte einer ordnungsgemäßen Analyse von Urheberrechtsverletzungen überspringt, die die Identifizierung der verantwortlichen Partei und das Verständnis des gesamten Prozesses von der Erstellung des Datensatzes bis zur Ausgabe der KI umfassen (Murray 2023; Lemley 2024). Die Frage, ob ein von einem KI-Modell erzeugtes Werk urheberrechtlich geschützt werden kann, ist momentan nicht klar, da KI keinen rechtlichen Urheberschaftsstatus hat und noch unklar ist, wem die Rechte an KI-generierten Inhalten gehören. Eine zu schnelle Fokussierung auf Verletzungen und Fair Use ohne diese Schritte schafft Verwirrung und übersieht die Komplexität, wer tatsächlich für mögliche Urheberrechtsverletzungen verantwortlich ist – die Datensatz-Ersteller:innen, die KI-Entwickler:innen oder die Endnutzer:innen, die den endgültigen Inhalt erstellen (Lee et al. 2023; Murray 2023).

Datenschutzprobleme bei KI-generierten Inhalten und den verwendeten Daten

Ein anderer wichtiger Aspekt ist der Datenschutz. Da KI-Modelle manchmal unter Verwendung sensibler oder personenbezogener Daten, einschließlich Daten von Minderjährigen, trainiert werden, wenn der Datensatz beispielsweise nicht ordnungsgemäß anonymisiert wurde (Wang et al. 2023). Dies kann zu Datenschutzverletzungen führen, wenn die von der KI generierten Ergebnisse persönliche Informationen ohne Zustimmung der betroffenen Personen zeigen (Baldassarre et al. 2023). Dies kann besonders kritisch sein, wenn es ermöglicht, Ein-

zelpersonen basierend auf den Ergebnissen des KI-Modells wieder zu identifizieren (Kollapally/Geller 2024). Es könnten verschiedene Probleme auftreten, von möglichen Überwachungsanwendungen und Verwundbarkeit gegenüber Cyberangriffen bis hin zu Profiling. Die Verwendung großer Datensätze mit sensiblen Daten zum Trainieren von KI-Modellen birgt Risiken aufgrund möglicher Verstöße und Datenschutzverletzungen, wie durch das italienische Verbot im Frühjahr 2023 hervorgehoben wurde, das OpenAI dazu veranlasste, seine Datenschutzrichtlinie zu erweitern und vor der Anmeldung der Nutzer zugänglich zu machen (Baldassarre et al. 2023; Tagesschau 2023). Es gibt auch ethische Überlegungen zur Verwendung von Daten für das Training von KI-Modellen. Zum Beispiel wirft die Verwendung von Daten ohne Zustimmung oder von Quellen, die möglicherweise nicht für diesen Zweck vorgesehen waren, ethische Fragen bezüglich des Respekts vor den Rechten der Einzelnen betroffenen Personen und der potenziellen Gefährdung auf (Golda et al. 2024).

Gesetzgebung nicht angepasst

Abschließend ist zu sagen, dass Generative KI weithin als bahnbrechende Technologie mit großem Potenzial zur Beschleunigung des Fortschritts in vielen Bereichen angesehen wird. Viele professionelle Künstler:innen, Schriftsteller:innen und Programmierer:innen sprechen sich jedoch entschieden gegen die Nutzung ihrer Werke als Trainingsdaten für KI-Systeme sowie gegen die Erstellung von Ergebnissen aus, die potenziell ihre Originalwerke ersetzen oder damit konkurrieren könnten. Die ursprünglichen Schöpfer:innen kritisieren auch das Fehlen von Anerkennung und Vergütung für die Nutzung ihrer Kunst (Samuelson 2023) und sie haben Probleme, ihre Werke zu schützen, da die aktuelle Gesetzgebung nicht auf diese Art von Situation angepasst ist (Lemley 2024). Die EU-Urheberrechtsrichtlinie[136] ist derzeit nicht in der Lage, alle Situationen im Zusammenhang mit Urheberrecht und Generativer KI abzudecken und hat nur einen relevanten Artikel dazu (Artikel 4 zu Ausnahmen und Beschränkungen für das Text und Data Mining) (Quintais 2024). Kürzlich hat der AI-Act zwei wesentliche Verpflichtungen für Anbieter von KI-Modellen im Zusammenhang mit generativer KI eingeführt, die diesen Artikel 4 ergänzen, nämlich Transparenzpflicht und Angabe der genutzten urheberrechtlich geschützten Werke. Die Wirksamkeit wird jedoch infrage gestellt, da der öffentliche Rechtsrahmen des AI-Act nicht gut mit der privaten Natur des Urheberrechts übereinstimmt. Es gibt zu viele Unterschiede in den Verpflichtungen, der Durchsetzung und den Rechtsmitteln. Dennoch könnten die Regelungen des AI-Act zum Urheberrecht dazu beitragen, dass KI-Modelle besser mit dem Urheberrecht übereinstimmen und möglicherweise auch indirekt zur Durchsetzung von Artikel 4 der EU-Urheberrechtsrichtlinie beitragen (Quintais 2024).

[136] Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG. ABl. EU L 130/92 vom 17.5.2019.

5.5 ZWISCHENFAZIT

Aus dem Vorstehenden ergibt sich der Befund, dass es – ganz unabhängig von den mit Generativer KI verbundenen Chancen und unabhängig von den spezifisch in diesem Bericht diskutierten Folgen für die Demokratie – gravierende Problemfelder gibt, die als Kontext zu jeder engeren Befassung mit einem Teilbereich (etwa den Auswirkungen auf die Demokratie) mitbedacht werden sollten. Besonders hervorzuheben sind

- der enorme Ressourcenverbrauch (Energie, Wasser, Land),
- der damit einhergehende bedeutsame CO_2-Fußabdruck,
- das gravierende ethische Defizit bei den Arbeitsbedingungen im menschlichen Teil der KI-Wertschöpfungskette,
- die weitgehend offenen Urheberrechts- und Datenschutzfragen sowie
- die weiteren sich abzeichnenden, aber noch wenig untersuchten Folgen für die Gesellschaft,
- insbesondere die disruptiven Auswirkungen auf den Bildungssektor.

Ohne eingehende Diskussion der genannten Wirkungen und ohne Setzen entsprechender Maßnahmen erscheint es aus TA-Perspektive nicht vertretbar, diese neuartige, sehr wirkmächtige Technologie einzusetzen, egal für welchen Zweck.

6 HANDLUNGSOPTIONEN

Aussage des OECD-Berichts steht für viele

> *„Governments will need to respond to the changes introduced by generative AI to the information space."* (Matasick et al. 2024, ch. 2.4.3)

Vor dem Hintergrund der Analyse der Chancen und Risiken Generativer KI in Hinblick auf das demokratische politische System stellt sich wie bei jeder neuen Technologie die Frage: Kann die Gesellschaft die angestoßene sozio-technische Entwicklung passiv beobachten und gewähren lassen? Oder ist es vielmehr angezeigt, den Prozess zu gestalten, d. h. in der einen oder anderen Weise zu versuchen einzugreifen? Dies kann durch Förderung erwünschter Entwicklungsrichtungen geschehen oder durch technische, organisatorische bzw. juristische Maßnahmen mit dem Ziel, Risiken auszuschließen oder zumindest zu minimieren.

Einheitliche Meinung der Beobachter:innen und Expert:innen:

Es besteht grundsätzlich Handlungsbedarf …

Die meisten Beobachter:innen und Expert:innen, die am Workshop im Rahmen dieser Studie (siehe Anhang) einbezogen oder deren hier verarbeitete Berichte und wissenschaftliche Literatur herangezogen wurden, kommen – wie der einleitend zitierte aktuelle OECD-Bericht – zum Schluss, dass grundsätzlich Handlungsbedarf besteht. Das Hauptargument ist, dass die vielfältigen Anwendungen dieser neuen Technik das Potenzial haben, wichtige Pfeiler der Demokratie zu erschüttern, nämlich die Gewährleistung eines rationalen Diskurses der Bürger:innen und politisch Aktiven sowie eine freie, d. h. ohne fremde Interventionen, insb. ohne Desinformation, zustande gekommene Meinungsbildung und Wahlentscheidung.[137]

… und es bedarf einer Kombination unterschiedlicher Maßnahmen

Weiters kommen viele Expert:innen zum Schluss, dass nur eine Kombination unterschiedlicher Maßnahmen den Herausforderungen gerecht werden kann. Das folgende Zitat steht für viele Einschätzungen:

> *„Die Kombination aus voreingenommenen (biased) Empfehlungsalgorithmen und gefälschten politischen Inhalten ist eine große Gefahr für die Demokratie. Um hier Abhilfe zu schaffen, brauchen wir eine Mischung aus technologischen Lösungen (zum Beispiel Anti-Deepfake-Detektoren), regulatorischen Lösungen (zum Beispiel Transparenzverpflichtungen für Algorithmen und Beschränkungen für die politische Nutzung von KI) und kulturellen Veränderungen beziehungsweise Wandel im Bildungsbereich (zum Beispiel Aufklärung der Bürgerinnen und Bürger, nicht jeden Inhalt für bare Münze zu nehmen und politische Inhalte mit größerer Sorgfalt zu prüfen)."* (Raymond Sun)[138]

In diesem Kapitel werden daher die verschiedenen Vorschläge für Reaktionen von Gesellschaft und Politik auf Generative KI speziell im Bereich Politik und Demokratie systematisiert und analysiert. Das Kapitel ist dreigeteilt und widmet sich zunächst möglichen regulativen Maßnahmen (6.1), anschließend Vorschlägen für organisatorische und institutionelle Ansätze (6.2) und schließlich möglichen technischen Methoden (6.3). Dieses Kapitel fußt auf Literatur- und Internetrecherche sowie auf den Ergebnissen des o. g. Workshops.

---

[137] Hier nur am Rande behandelt werden Bedrohungen der öffentlichen Ordnung, die durch Hass und Verhetzung in den Sozialen Medien entstehen, da dies für demokratische wie auch nicht-demokratische Systeme gleichermaßen gilt, wobei sich demokratische Rechtsstaaten naturgemäß schwerer tun, mäßigend einzugreifen – was ebenfalls ein gravierendes Problem für demokratische Systeme ist.

[138] In human 2023/2, S. 41.

6.1 REGULATIVE ANSÄTZE

Scoping dieses Abschnitts

In diesem Abschnitt wird „Regulierung" in einem weiten Sinne verstanden und bezieht alle rechtlichen Formen mit ein (soft und hard law). Es sei weiters vorausgeschickt, dass im Rahmen dieser Studie weder eine Exegese der geltenden Rechtlage noch eine detaillierte Analyse der rechtlichen Optionen erfolgen kann – dies bleibt juristischen Expert:innen vorbehalten. Vielmehr werden die in Rechtsakten bzw. Vorschlägen zu solchen, aber auch von Expert:innen und Beobachter:innen ins Spiel gebrachten regulativen Optionen systematisch dargestellt.

6.1.1 DIE RECHTSLAGE IM ÜBERBLICK

(Generative) KI ist nicht (mehr) unreguliert

Auch wenn KI und insbesondere Generative KI noch relativ neue Techniken darstellen, gibt es bereits zahlreiche rechtliche Vorschriften, die auf unterschiedliche Aspekte ihrer Anwendung Bezug nehmen. Dabei kann man eine Reihe unterschiedlicher rechtssetzender Akteure unterscheiden:

Rechtssetzende Akteure

- *Internationale Organisationen* (UNO etc.) bereiten völkerrechtliche Verträge vor, die dann national oder supranational ratifiziert und umgesetzt werden.
- Die *Europäische Union*, also die EU-Kommission als Vorschlagende sowie Ministerrat und Europäisches Parlament, agieren als Gesetzgeber/Legislative für sog. supranationale Rechtsakte und bestimmen damit den Rechtsrahmen für den Umgang und Einsatz neuer Techniken wie der KI, auch der Generativen KI.
- Die *nationalen Parlamente* sind als Legislative kompetent (im Rahmen des EU-rechtlich Zulässigen) Gesetze für das eigene Territorium zu verabschieden. Gerade im Bereich des Internets und der Informationstechnologien ist die Durchsetzung solcher Gesetze im nationalstaatlichen Rahmen freilich schwierig (siehe Abschnitt 5.1 zur Digitalen Souveränität). Bisweilen agieren nationale Gesetzgeber schneller als die EU-Ebene, müssen aber nach EU-Beschlussfassung das nationale Gesetz anpassen oder sogar abschaffen.
- *Alle Parlamente* sind weiters kompetent, für eigene Aktivitäten über ihre Geschäftsordnungen (Gesetze) Regeln zu setzen, etwa ob der Parlamentsbetrieb durch Anwendungen auf Basis Generativer KI unterstützt werden soll.
- Die *nationalen Regierungen* wie auch die *EU-Kommission* agieren nicht nur als Initiatoren der Gesetzgebung, sondern auch als politische Akteure, die Aktionspläne, Programme, Mitteilungen, Guidelines usw. veröffentlichen, die dann als Soft-Law das Geschehen mitbeeinflussen können.
- Die (wahlwerbenden) *politischen Parteien* haben eine große Verantwortung, den politischen Diskurs konstruktiv zu gestalten und könnten, entweder alleine oder in Absprache mit anderen Parteien, ihr Handeln regeln, also etwa den Einsatz von Generativer KI im Wahlkampf beschränken (Verhaltenskodizes).
- Auch die *Plattform-Provider* und *Tech-Unternehmen* können sich selbst und ihren Kund:innen (in AGBs, Richtlinien) oder ihrer Branche (Standards, Codes of Conduct) Regeln geben, wie Generative KI produziert und eingesetzt wird. Die folgende Tabelle 4 gibt einen Überblick über die angesprochenen Regulierungsarten und ihre Akteure, ergänzt um je ein bis zwei Beispiele.

Tabelle 4: Überblick über Regulierungsarten

Regulierungsarten	Beispiele	Akteure
HARD LAW		
Nationales Gesetz	TelekommunikationsG DSA-Begleitgesetz	Österr. Parlament
Supranationaler Rechtsakt	VO Transparenz–Targeting–politische Werbung; AI-Act	Europ. Parlament und Rat
Völkerrechtlicher Vertrag	Div. Urheberrechts-Abkommen	UNO-WIPO
Akt einer Behörde	Div. Verordnungen über Lehrpläne	BMBWK
Geschäftsordnung	GeschäftsordnungsG	Österr. Parlament
SOFT LAW: VERHALTENSKODICES (CODES OF CONDUCT)		
Selbstregulierung durch Akteure	KI-Kodex 2023[139]; Richtlinie Falschmeldungen[140]	Schweizer Mitte-Links-Parteien; Facebook
Ko-Regulierung (Behörden plus Akteure)	Strengthened Code of Practice; on Disinformation 2022[141]	EU-Kommission und Industriepartner

Regulative Aktivitäten auf EU-Ebene

Auf EU-Ebene gibt es mittlerweile ein ausdifferenziertes und immer komplexer werdendes Normengeflecht, das Aktivitäten im Bereich IT im Allgemeinen und KI im Besonderen zu regeln versucht. Zu nennen sind insbesondere der DSA, also jener Rechtsakt, der sich mit der Erbringung digitaler Dienstleistungen befasst, die DSGVO, die die Basis für den Datenschutz in Europa bildet, der AI-Act, der den Einsatz Künstlicher Intelligenz regelt, und speziell die Verordnung über Transparenz und Targeting politischer Werbung sowie einige weitere.[142] Hervorzuheben ist, dass diese Rechtsakte als sog. Verordnungen, nicht als Richtlinien konzipiert wurden, also direkt in Österreich im Gesetzesrang anwendbar sind und nicht erst innerstaatlich umgesetzt werden müssen. Generative KI kommt in den EU-Rechtsakten außer im AI-Act, nur indirekt vor. (Generative) KI im Zusammenhang mit demokratischen Prozessen gilt dabei als Hochrisikoanwendung, für die strengere Vorschriften gemäß AI-Act gelten:

> *„KI-Systeme, die bestimmungsgemäß verwendet werden sollen, um das Ergebnis einer Wahl oder eines Referendums oder das Wahlverhalten natürlicher Personen bei der Ausübung ihres Wahlrechts bei einer Wahl oder einem Referendum zu beeinflussen.“* (Anhang 3, Ziffer 8 (b) AI-Act)

139 *netzwoche.ch/news/2023-09-26/mitte-links-parteien-verpflichten-sich-zur-ki-deklaration*.

140 *transparency.meta.com/de-de/policies/community-standards/misinformation/*.

141 *disinfocode.eu*.

142 Es ist nicht einfach, in diesem dynamischen Rechtsbereich einen aktuellen Überblick über die Rechtslage auf EU-Ebene zu geben. Siehe dazu etwa die beeindruckende Übersicht des Think Tanks Bruegel, *bruegel.org/sites/default/files/private/2023-11/Bruegel_factsheet.pdf*, welches die über 100 (!) EU-Rechtsakte im Digitalbereich darstellt. Für den Bereich der Regeln für KI und Privatsphäre, die insbesondere Konsument:innen betreffen, siehe beispielsweise Krieger-Lamina/Peissl (2024, im Erscheinen, S. 36ff). Zur EU-Regulierung von Deepfakes siehe den Überblick in Karaboga et al. (2024, S. 204ff); zum DSA siehe Matasick et al. (2024, ch. 2.2).

Was bringt die EU-Regulierung?

All diese Rechtsakte sind in den üblichen langwierigen Verhandlungsprozessen auf EU-Ebene unter massivem Einfluss der großen Interessensvertretungen zustande gekommen. Da sonst noch praktisch kaum ein Gesetzgeber weltweit überhaupt Regeln in diesem Bereich verabschiedet hat,[143] haben die regulativen Aktivitäten der EU weltweit Vorreitercharakter. Nach Ansicht vieler Beobachter:innen dauern die Rechtssetzungsverfahren angesichts der Dynamik bei der Technik- und Marktentwicklung zu lange und gehen die EU-Regulierungen nicht weit genug, um die adressierten Risiken effektiv zu entschärfen. Jurist:innen und Politolog:innen untersuchen daher bereits intensiv, ob die EU-Regelungen konsistent sind und den an sie gestellten Anforderungen genügen (siehe bspw. Casero-Ripollés et al. 2023; Karaboga 2023; Cupać/Sienknecht 2024; Łabuz 2024). Darüber hinaus zeigt die Erfahrung, etwa mit der DSGVO, dass die betroffenen Unternehmen, insbesondere „BigTech", alles daransetzen, ihre Aktivitäten auch am Rande oder jenseits der Legalität fortzusetzen und sich gegebenenfalls klagen zu lassen, wenn es kommerziell opportun erscheint.[144] Die darauffolgenden Verfahren (etwa vor dem EuGH) dauern in der Regel jahrelang, währenddessen der Betrieb ungestört weiterläuft; selbst wenn Strafen ausgesprochen werden, werden diese angefochten oder sind aus kommerzieller Sicht angesichts hoher Gewinne offenbar nicht ausreichend abschreckend (ausführlich dazu Zuboff 2019)[145] (siehe dazu schon Abschnitt 4.3 zur Machtkonzentration).

Rechtslage in Österreich

In Österreich[146] gibt es bislang praktisch keine nationalen, speziell auf KI bzw. Generative KI anwendbare Vorschriften, vielmehr gelten neben dem (direkt anwendbaren[147]) EU-Recht allgemeine Vorschriften aus den Bereichen Telekommunikation, Datenschutz, Urheberrecht, Wettbewerbsrecht, Strafrecht usw. die auch bei (generativer) KI relevant sind. Darüber hinaus waren zur Flankierung der Implementierung der EU-Regelungen vereinzelt Begleitmaßnahmen notwendig. So implementiert das DSA-Begleitgesetz[148] das EU-Gesetz über digitale Dienste in Österreich: Es wird ein Koordinator für Digitale Dienste (KDD) eingerichtet (KommAustria), der die Einhaltung des DSA für Vermittlungsdienste mit Sitz in Österreich überwacht. Darüber hinaus wird dort geregelt, welche Behörden und Gerichte in Österreich an der Überwachung und Regulierung solcher Online-Dienste beteiligt sind.[149] Zur begleitenden Umsetzung des AI-Acts wurde die in

---

143 Eine Ausnahme ist Kalifornien, siehe die eindrucksvolle Liste an Gesetzen, die hier kurz beschrieben werden: ***gov.ca.gov/2024/09/29/governor-newsom-announces-new-initiatives-to-advance-safe-and-responsible-ai-protect-californians/***.

144 Ein rezentes Beispiel betrifft den Mutterkonzern von Facebook, Meta: „*Die EU-Kommission verdächtigt Meta, zu wenig gegen die Verbreitung von Falschinformationen auf seinen Plattformen Facebook und Instagram zu tun, und hat ein Verfahren [nach dem DSA, bereits das fünfte] gegen den US-Konzern eingeleitet. Konkret untersucht werden soll, ob Meta im Umgang mit politischer Werbung europäische Regeln verletzt.*" ***orf.at/stories/3355952/***.

145 Ob die zweistelligen Milliardenstrafen für Apple und Google, die der Europäische Gerichtshof jüngst bestätigt hat, diese Konzerne zu zukünftig rechtskonformen Verhalten anhalten werden, bleibt abzuwarten (***orf.at/stories/3369099/***).

146 Zur Rechtslage in der Schweiz, speziell zu Deepfakes, siehe Karaboga et al. (2024, S. 159ff).

147 Die Bestimmungen des AI-Act treten zeitlich abgestuft in Kraft, siehe ***rtr.at/rtr/service/ki-servicestelle/ai-act/Zeitplan.de.html***.
Die Bestimmungen über „general purpose AI" treten Anfang August 2025 in Kraft.

148 BGBl. 182/2023.

149 Siehe auch ***bmj.gv.at/themen/EU-und-Internationales/Digital-Services-Act.html***.

der RTR eingerichtete Servicestelle für Künstliche Intelligenz als Ansprechpartner und Informationshub für die breite Öffentlichkeit zum Thema KI eingerichtet.[150] Zur Rechtslage in Hinblick auf Deepfakes siehe auch den Aktionsplan Deepfakes.[151]

6.1.2 REGULIERUNGSINHALTE UND -OPTIONEN IM ÜBERBLICK

Wie einleitend klargestellt, kann im Rahmen dieser Studie kein detaillierter und umfassender Überblick über die komplexe Rechtslage zum Thema Generative KI gegeben werden. Was hier jedoch geleistet werden soll, ist die Systematisierung potenzieller (also vorgeschlagener, diskutierter bzw. teils beschlossener) Regulierungsinhalte bzw. -optionen, siehe die nachfolgende Tabelle 5. Prinzipiell sind die Regulierungsoptionen anhand des Lebenszyklus von Generativen KI-Anwendungen geordnet, von der Produktion, über den allgemeinen Einsatz bis zum Einsatz in der Politik. Optionen zur Durchsetzung dieser möglichen regulatorischen Maßnahmen schließen die Tabelle ab.

Fünf Regulierungstypen

Anhand der gefundenen Regulierungsoptionen und auf Basis eines Analogieschlusses von anderen Technologieregulierungen können mögliche Typen von Regulierung unterschieden werden, die auf unterschiedlichen Ebenen und mit verschiedenen Ansätzen versuchen, die Erstellung, Verbreitung und Verwendung von Generativer KI lenkend einzugreifen (siehe linke Spalte in Tabelle 5). Hierbei konnten fünf übergeordnete Typen gebildet werden, die als Heuristik für die Einordnung von Regulierungsoptionen verwendet werden können (siehe Abbildung 5).

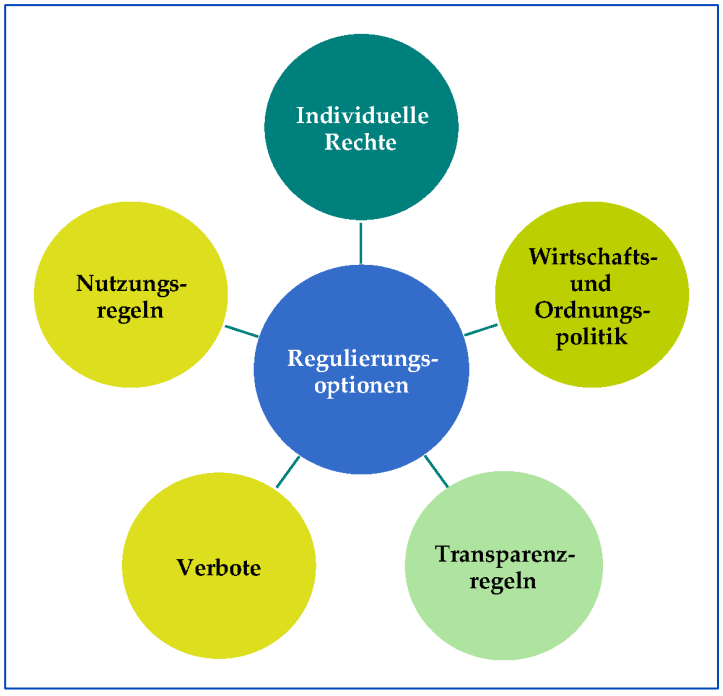


Abbildung 5: Schematische Darstellung (denkbarer) Regulierungsoptionen

150 *rtr.at/rtr/service/ki-servicestelle/ki-servicestelle.de.html*.
151 *bmi.gv.at/bmi_documents/2779.pdf*, S. 14ff.

Quelle: Eigene Darstellung

Individuelle Rechte

Zum einen besteht ein Regulierungsansatz über die Definition und Ausgestaltung von Einforderungsmöglichkeiten individueller Rechte. So können Rechte über die informationelle Selbstbestimmung auch die Verwendung von personenbezogenen Daten für das Training von Generativer KI in Frage stellen. Ähnlich lässt sich dies bei Eigentumsrechten an z. B. immateriellen Gütern festhalten. Ebensolche Rechte wären dann z. B. in zivilgerichtlichen Verfahren gegenüber z. B. Erstellern oder Nutzer:innen von Anwendungen Generativer KI einklagbar.

Nutzungsvorschriften und Anpassung bestehender Regeln

Ein weiteres großes Thema lässt sich unter dem Überbegriff Nutzungsvorschriften zusammenfassen, wobei hier auch die Anpassung bestehender Regeln genannt werden kann. Es kann sich in diesem Fall um Ordnungswidrigkeiten handeln, die z. B. Regelverstöße von Plattformbetreibern sanktionieren und eher dem Verwaltungsrecht zugeordnet werden können. Die Sanktionen können hierbei von Bußgeldern bis hin zum Verbot von (wirtschaftlichen) Tätigkeiten reichen. Auch Kodifizierungen im (Neben-)Strafrecht sind hier zu bedenken (bspw. durch das VerbandsverantwortlichkeitsG).

Verbote

Eng damit zusammenhängend als Regulierungsoption sind Verbote, die bestimmte Handlungen prinzipiell untersagen, wie z. B. das Verbot von Microtargeting ohne vorherige Zustimmung (wie in der Targeting-VO) oder ein mögliches Verbot der Verbreitung von Deepfakes von Politiker:innen. Wie auch schon bei den Nutzungsvorschriften können Verbote wiederum als Ordnungswidrigkeit oder Straftat reguliert werden.

Transparenzregeln

Ein zentraler Regulierungsansatz liegt in speziellen Vorschriften zur Kennzeichnung und Sichtbarmachung von durch Generative KI generierten Inhalten, entweder menschlich sichtbar oder für Maschinen erkennbar, wodurch z. B. Browser Hinweise, dass es sich um KI-Generiertes handelt, für Endnutzer:innen setzen können. Durch eine Pflicht zur Kennzeichnung würde bis zu einem gewissen Ausmaß Transparenz hergestellt werden. Erst durch diese transparente Darstellung von durch Generative KI erstellten Inhalten würde eine wirkliche Einordnung von Generativer KI durch Nutzer:innen möglich werden, um informierte Entscheidungen bezüglich Konsum und (politischer) Meinungsbildung treffen zu können.

Wirtschafts- und Ordnungspolitik

Weitere vieldiskutierte Regulierungsansätze liegen im Bereich der Wirtschafts- und Ordnungspolitik und haben mit der Bekämpfung von Machtkonzentration und der Schaffung eigener nationaler bzw. europäischer Champions zu tun. Die Bekämpfung von Monopolen oder anderen Arten von Marktverzerrung im Bereich Generativer KI kann so mit klassischer Ordnungspolitik wie dem Kartellrecht bekämpft reguliert werden. Die Unterstützung von neuen Startups oder bestehenden heimischen Unternehmen kann so mit Instrumenten der Wirtschaftsförderung wie Fördergeldern oder bevorzugter Besteuerung angegangen werden.

Tabelle 5: Systematik von Regulierungsoptionen für Generative KI

Regulierungstyp	Regelungsinhalt	Beschreibung, Anmerkungen	Wo geregelt bzw. vorgeschlagen
PRODUKTION VON GENERATIVER KI			
Indiv. Rechte	Urheberschutz	Urheber:innen von Texten, Bildern etc., die zum Trainieren der KI verwendet werden, müssen dieser Verwendung zustimmen	Siehe EU-Urheberrechtsrichtlinie div. Erw. (z. B. 105) und Art. 53 AI-Act
Wirtsch./ Ordnungspolitik	Kartellrecht	Entflechtungen, um Monopolbildung zu vermeiden; Akquisitionskontrollen	Siehe EU-Kartellrecht
Transparenz	Kennzeichnung: Maschinenlesbare Wasserzeichen	Generative-KI-Software muss prinzipiell maschinenlesbare Wasserzeichen hinterlegen, die von Browsern etc. ausgelesen werden können, um sie den Nutzer:innen anzuzeigen	Art. 50 (2) AI-Act Deepfakes Accountability Act der USA[152] Partnership on AI's (PAI) Responsible Practices for Synthetic Media[153]
Indiv. Rechte	Datenschutz	In den Trainingsdaten dürfen keine personenbezogenen Daten ohne explizite Zustimmung der Betroffenen enthalten sein	u. a. Erw. 69 AI-Act Siehe auch DSGVO
Wirtsch./ Ordnungspolitik	Vermeidung von Bias	Anwendungen Generativer KI sollen Menschen nicht diskriminieren, also nach Geschlecht, Hautfarbe, Religion etc. unterschiedlich behandeln	insb. Erw. 27 AI-Act
Transparenz	Offenlegung der Trainingsdaten	Verpflichtung zur detaillierten Dokumentation der Datensätze, die bei der Entwicklung und Schulung der KI-Technologie oder -Dienstleistung verwendet wurden	AB 2013/Kalifornien[154]
EINSATZ VON GENERATIVER KI			
Nutzungsregel	Einhaltung von Standards, Konformitätsprüfungen für Hochrisikoanwendungen	Für den Einsatz von KI-Hochrisikoanwendungen (zu denen auch Anwendungen im Bereich demokratischer Prozesse gehören) gelten spezifische, strenge Regelungen	Art. 43 etc. AI-Act
Wirtsch./ Ordnungspolitik	Cybersicherheit	Ein angemessenes Maß an Cybersicherheit für die KI-Modelle und die physische Infrastruktur des Modells sind zu gewährleisten.	Art. 55 AI-Act
Wirtsch./ Ordnungspolitik	Energieeffizienz, Ressourcenschonung und Nachhaltigkeit	Serverinfrastrukturen zum Betreiben Generativer KI sollten energieeffizient und ressourcenschonend (Wasser, Fläche) sein	[bislang nirgendwo]

[152] Gesetz US HR3230 "Deep Fakes Accountability Act Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019", *billtrack50.com/billdetail/1132741*.

[153] *syntheticmedia.partnershiponai.org*.

[154] Gesetz AB 2013 "Generative artificial intelligence: training data transparency" *leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013*.

Regulierungstyp	Regelungsinhalt	Beschreibung, Anmerkungen	Wo geregelt bzw. vorgeschlagen
Wirtsch./ Ordnungspolitik	Vermeidung systemischer Risiken	Durchführung von Modellbewertungen mit standardisierten Protokollen und Instrumenten, die dem Stand der Technik entsprechen	Erw. 99 und 110, Art. 51 und 55 AI-Act
Transparenz	Kennzeichnung: Für Nutzer:innen sichtbare Labels	Bei Verwendung von KI-generierten Inhalten ist das auszuweisen – ähnlich Kennzeichnung von Werbung als Anzeigen, Produktplatzierungen o. ä.	Art. 50 (2) AI-Act Art. 35 (1) k DSA
Transparenz	Transparenzregeln	Technische Dokumentation, Trainingsdaten	Art. 11 (1) AI-Act
Verbot	Generelles Verbot zur Verbreitung von Desinformation	Auf Basis einer generellen Definition von Desinformation wird ein Verfahren aufgesetzt, dass dieses feststellen kann	[bislang nirgendwo]
Verbot	Verbot von Bots, die sich als Mensch ausgeben	Kein Verbot von Bots generell, jedoch sollen sich nicht-menschliche Akteure, die durch Generative KI gesteuert werden, ihre Kommunikationspartner nicht im Unklaren über ihre Künstlichkeit lassen	Harari (2024, S. 473f.) Art. 50 (1) AI-Act
GENERATIVE KI SPEZIELL IN DER POLITIK			
Nutzungsregel	Verbindlichmachen von Code of Conducts	Z. B. des „Code of Practice on Disinformation", inklusive regelmäßiger externer Audits	Villar García et al. (2021, S. 101f)
Nutzungsregel	Verhaltenskodex für Politiker:innen im Umgang mit Generativer KI	Dieser Kodex enthält eine Reihe von Regeln über den Gebrauch von Generativer KI im politischen Geschäft, inklusive Wahlkampf	Z. B. KI-Kodex 2023 der Schweizer Mitte-Links-Parteien[139]
Verbot	Budgetbeschränkung für den Einsatz von Generativer KI im Wahlkampf	Um zu verhindern, dass der Wahlkampf nur mehr durch KI anstatt durch Menschen geführt wird.	Kruschinski (2023 ab Minute 38)
Verbot	Verbot von Microtargeting im politischen Kontext ohne Zustimmung der Adressierten	Microtargeting ist nur mit Zustimmung zulässig	Art. 18 VO Targeting[155]
Verbot	Verbot von Microtargeting im politischen Kontext	Generelles Verbot von Microtargeting	[bislang nirgendwo]
Transparenz	Kennzeichnung von politischen Anzeigen	Politische Werbung auf Plattformen muss als solche inkl. Urheberschaft gekennzeichnet werden (generell, gilt aber auch für KI-generiertes Microtargeting)	Art. 11f VO Targeting
Verbot	Kein Profiling und Microtargeting von Nicht-Wahlberechtigten	Politische Werbung darf nicht gezielt an Personen bis ein Jahr vor Erreichung des Wahlalters verschickt werden	Art. 18 Abs 2 VO Targeting

[155] *eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202400900*; diese Verordnung tritt ab dem 10. Oktober 2025 in Kraft.

Regulierungstyp	Regelungsinhalt	Beschreibung, Anmerkungen	Wo geregelt bzw. vorgeschlagen
Verbot	Verbot von Deepfakes von Politiker:innen	Generelle Einschränkung der Nutzung von Deepfakes (für Negativkampagnen) mit Bezug zu Politiker:innen	KI-Kodex 2023 der Schweizer Mitte-Links-Parteien[139]
Verbot	Verbot von Deepfakes politischer Kandidat:innen, innerhalb 60 Tagen vor einer Wahl, die den Betroffenen schaden sollen	Zeitliche Einschränkung des Verbots; [satirisches Deepfakes, also ohne Schädigungsabsicht, blieben zulässig]	AB 2839/Kalifornien[156]
Nutzungsregel	Digitaler Ordnungsruf	Bei Missachtung eines – noch zu beschließenden – Verhaltenskodex für Umgang mit Sozialen Medien (ähnlich Ordnungsrufen im Parlamentsplenum durch Parlamentspräsidenten/in)	Karmasin et al. (2024, S. 35)
Verbot	Verbot von Algorithmen, die unüberwacht wichtige öffentliche Debatten kuratieren	Algorithmen sollten nie alleine entscheiden, welche Stimmen sie zum Schweigen bringen und welche sie verstärken	Harari (2024, S. 475)
Nutzungsregel	Blockieren irreführender wahlbezogener Inhalte	Bekämpfung von Online-Desinformation durch Verpflichtung großer Plattformen, irreführende wahlbezogene Inhalte während bestimmter Zeiträume vor und nach einer Wahl zu blockieren oder zu kennzeichnen	AB 2655/Kalifornien[157]
		RECHTSDURCHSETZUNG	
Wirtsch./ Ordnungspolitik	Kontrollbehörden	Eigene Behörden sind als Drehscheibe für die Information der Bürger:innen und Kontrolle des Einsatzes von (Generativer) KI einzurichten	AI-Act (Kap. 7), Art. 22 VO Transparenz
Nutzungsregel	Inhaltliche Verantwortlichkeit der Plattformbetreiber	(Alle) Plattformbetreiber müssen *aktiv* nach rechtswidrigen Inhalten und/oder Desinformation suchen und diese löschen (nicht nur auf behördliche Anordnung)	[noch nirgendwo]
Nutzungsregel	Verpflichtung der Plattformbetreiber zur Löschung von bestimmten Inhalten	Plattformbetreiber müssen löschen, wenn sie von rechtswidrigen Inhalten und/oder Desinformation erfahren	Art. 6 DSA Karaboga et al. (2024, S. 333)
Indiv. Rechte	Meldesysteme	Meldeverfahren für Inhalte, Abhilfeverfahren inkl. Wahrung der Rechte der von Meldungen betroffenen Nutzer:innen, internes Beschwerdemanagementsystem	Art. 16 und 20 DSA
Indiv. Rechte	Beschleunigte Verfahren	Prioritäre Behandlung von Meldungen über illegalen Content durch vertrauenswürdige Hinweisgeber	Art. 22 (1) DSA

156 Gesetz AB 2839 "Elections: deceptive media in advertisements" *leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2839*.

157 Gesetz AB 2655 "Defending Democracy from Deepfake Deception Act of 2024", *leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2655*.

Regulierungstyp	Regelungsinhalt	Beschreibung, Anmerkungen	Wo geregelt bzw. vorgeschlagen
Transparenz	Risikobewertung und Berichterstattung	Sehr große Plattformen/Suchmaschinen ermitteln, analysieren und bewerten sorgfältig alle systemischen Risiken und erstatten über Maßnahmen zur Risikominimierung jährlich Bericht	Abschnitt 5 DSA Kap. 9 AI-Act
Nutzungsregel	Regelung von digitalen Beweisen im Strafverfahrensrecht	Deepfakes zwecks Visualisierung von Tathergängen oder zur virtuellen Tatortbegehung	Karaboga et al. (2024, S. 333)

6.1.3 DISKUSSION AUSGEWÄHLTER REGULATIVER OPTIONEN

Wie in Tabelle 5 ersichtlich, wurden in Europa, in den USA, insb. Kalifornien und in anderen Staaten[158] bereits einige Regulierungsmaßnahmen im Zusammenhang mit Generativer KI generell und der politischen Sphäre im Besonderen gesetzt. Diese Maßnahmen werden hier weder dargestellt noch aus juristischer Sicht analysiert. In diesem Abschnitt werden vielmehr Vorschläge für neue bzw. Weiterentwicklungen bestehender Regulierungen hervorgehoben, über die es also noch politische Debatten braucht.

Plattformen gleichsam als Medien behandeln

Die aktuelle Rechtslage stuft die Betreiber der Plattformen, auf denen ein größer werdender Anteil des politischen Diskurses stattfindet, nicht als Medien, sondern als Digitale Dienstleister ein. Dies hat unter anderem zur Folge, dass diese anders als herkömmliche Medien, inklusive deren digitalen Varianten, nicht für die auf den Plattformen veröffentlichten Inhalte verantwortlich sind. Von vielen Seiten wird eine Änderung der Rechtslage dahingehend gefordert, die inhaltliche (Mit-)Verantwortlichkeit der Plattformbetreiber festzustellen, sie also gleichsam als Medien einzustufen. Das würde mit der Verpflichtung einhergehen, aktiv nach rechtswidrigen Inhalten und/oder Desinformation zu suchen und diese zu löschen. Mit dieser Option müsste teilweise rechtliches Neuland betreten werden bzw. an der Urheberrechtsgesetzgebung[159] Anlehnung genommen werden; eine bloße Übertragung des Medienrechts, das ja für vergleichsweise kleine Medien mit überschaubarem Output per Tag und Redaktionen entwickelt wurde, erscheint angesichts der enormen täglichen Menge an user-generierten Inhalten kaum vorstellbar. Dementsprechend wird das Thema auch kontrovers diskutiert.[160] Dabei pochen die Plattformen darauf, nur als Intermediäre aufzutreten (obwohl sie großen Einfluss auf die Zusammenstellung, Reihung usw. der auf ihnen geposteten Nachrichten nehmen), während die traditionellen Medienunternehmen eine ihnen vergleichbare Verantwortlichkeit ihrer neuen, starken Mitbewerber einfordern.

---

158 Selbst in China, siehe: *reconnect-china.ugent.be/2024/10/30/reconnect-china-policy-brief-15-ai-and-technical-standardization-in-china-and-the-eu/*; *ferner-alsdorf.de/ueberblick-ueber-die-regulierung-von-kuenstlicher-intelligenz-ki-in-china/*; *china-briefing.com/news/chinas-neue-vorschriften-fur-generative-ki-verstehen/*; *table.media/china/sinolytics-radar/warum-auslaendische-ki-unternehmen-in-china-unter-druck-stehen/*.

159 Siehe Uploadfilter: *datenschutz-generator.de/ratgeber-urheberrecht-uploadfilter/*.

160 Siehe stellvertretend die Debatte im österr. Parlament, *ots.at/presseaussendung/OTS_20241112_OTS0151/demokratie-enquete-expertinnen-und-abgeordnete-loten-handlungsbedarf-in-oesterreich-aus*.

Tatsächlich handelt es sich um eine Mammutaufgabe, den enormen Traffic auf den Sozialen Plattformen laufend nach rechtswidrigen Inhalten zu durchforsten; diese kann vermutlich ebenfalls nur (teil-)automatisiert und damit unter maßgeblicher Beteiligung der Plattformen selbst geleistet werden. Auf dem Weg zu einer demokratieverträglichen Regelung müsste – ähnlich wie im Fall der urheberrechtsdurchsetzenden Uploadfilter – eine Balance zwischen Meinungsfreiheit, Vermeidung von (willkürlicher) Zensur, Schutz der User:innen vor verbalen Übergriffen und Demokratieverträglichkeit gefunden werden (Karaboga et al. 2024, S. 160ff; Karmasin et al. 2024, S. 31ff). Es ist zu erwarten, dass, in den Worten von Saxer (2023), das „herkömmliche Medienrecht […] als Regulierungskonzept ein Auslaufmodell" (S. 157) ist und die Entwicklung von „der national geprägte[n] Medien- zur umfassenden Kommunikationspolitik, vom analogen, nationalen Medien- zum digitalen, teilweise übernationalen Kommunikationsrecht" (S. 160) führen wird.

Kartellrecht konsequent zur Anwendung bringen

Im Zusammenhang mit der Rolle der Plattformen wird weiters gefordert, das Kartellrecht konsequent zu nutzen, um noch stärkere Machtkonzentration durch Generative KI bei wenigen internationalen Digitalunternehmen zu unterbinden. Die diesbezüglichen Aktivitäten der EU-Kommission müssten forciert und wohl in behördlicher Kooperation mit Wettbewerbsbehörden in den USA und anderen Ländern umgesetzt werden. Das kann bis zur Entflechtung von bereits bestehenden Monopolen/Oligopolen führen. Zur konsequenten Anwendung würde auch gehören, die entsprechenden behördlichen Kapazitäten aufzustocken, um rascher reagieren zu können, damit die Verfahren zur Rechtsdurchsetzung nicht zu lange dauern. Das gilt nicht nur für das Kartellrecht, sondern für alle Verstöße gegen die mühsam errungenen rechtlichen Bestimmungen.

Diverse Vorschläge für Verbote zum Schutz des politischen Diskurses

Ein weiteres Bündel regulativer Optionen betrifft verschiedene Verbote, die den politischen Diskurs vor Desinformation schützen sollen. So wurde ein *generelles Desinformationsverbot* in allen Medien, Sozialen Medien, Presseaussendungen etc. vorgeschlagen. Unabhängig davon, wie dieses durchgesetzt werden könnte, wäre möglicherweise allein das Österreich- bzw. EU-weite Aufstellen und Kommunizieren einer solchen Regel für manche Akteure abschreckend. Insgesamt wäre es wohl als Anstoß für eine breite Debatte über die Frage, was als Desinformation gelten soll. Bislang gibt es dazu ja keinen breiten Konsens, der aber im Zuge seriöser Analysen und Auseinandersetzungen schrittweise hergestellt werden könnte. Ein Kristallisationskern für diese notwendige Debatte könnte ein solches generelles Desinformationsverbot mit einer allgemeinen Legaldefinition sein. Weniger weitgehend, dafür verhältnismäßig leichter durchzusetzen wäre ein *Verbot von Deepfakes politischer Kandidat:innen*, welche den Betroffenen schaden sollen – bzw., allgemeiner formuliert, ein Verbot von „dirty campaigning". Dazu gibt es bereits verschiedene Ansätze, etwa dass das Verbot generell gilt (wie im Pakt der Schweizer Mitte-Links-Parteien 2023[139]) oder beispielsweise nur innerhalb der letzten 60 Tagen vor einer Wahl (wie z. B. in Kalifornien[157]). In der EU ist Microtargeting in der Politik zulässig, sofern von den Adressat:innen zugestimmt wurde (siehe Art. 18 VO Targeting[155]); verboten ist jedoch, politische Werbung gezielt an Personen bis ein Jahr vor Erreichung des Wahlalters zu verschicken. Da die praktische Umsetzung und der Nachweis, ob zugestimmt wurde, schwierig ist, wird auch vorgeschlagen, ein *generelles Verbot von Microtargeting zu*

politischen Zwecken zu normieren.[161] Ebenfalls ventiliert wurde ein *Verbot von Bots, die sich als Mensch ausgeben*; es geht also um kein generelles Verbot, jedoch sollen sich nicht-menschliche Akteure, die durch Generative KI gesteuert werden, ihre Kommunikationspartner:innen nicht im Unklaren über ihre Künstlichkeit lassen (Harari 2024, S. 473f.). Harari schlägt weiters vor, Algorithmen zu verbieten, die unüberwacht wichtige öffentliche Debatten kuratieren/moderieren: Algorithmen sollten nie alleine entscheiden, welche Stimmen sie zum Schweigen bringen und welche sie verstärken (S. 475).

Umgang mit desinformierenden Quellen

Wie viele andere auch, plädieren Matz et al. (2024) dafür, den Fokus der Bekämpfung von Desinformation weg von der Empfänger- hin zur Erzeugerseite zu verlegen, also den Plattformen und diejenigen, die Inhalte veröffentlichen, in die Pflicht zu nehmen. Weniger intrusiv als Verbote wäre dementsprechend etwa die generelle Vorschrift bzw. Empfehlung an Plattformen und Suchmaschinen, dass Einträge, Webseiten und Postings aller Arten, die von Internetquellen stammen, die nachweislich bereits vielfach Desinformation oder generierte Inhalte verbreitet haben, *in den Suchergebnissen bzw. Timelines nachgereiht* werden sollen. Positiv formuliert liefe das darauf hinaus, dass demokratieförderliche Inhalte priorisiert werden sollten. Dies wäre das gelindere Mittel gegenüber einem totalen Blocken von irreführenden, wahlbezogenen Informationen, was bereits in Kalifornien geltendes Recht ist[157] bzw. in Hinblick auf beispielsweise pornographische Inhalte bereits laufend gemacht wird.

Umfassende Kennzeichnung KI-generierter Inhalte und deren Durchsetzung

Die Kennzeichnung KI-generierter Inhalte ist bereits Gegenstand verschiedener Rechtsakte. Art. 50 (2) AI-Act verpflichtet die Unternehmen, die Tools für Generative KI auf den Markt bringen, die Ergebnisse in maschinenlesbarer Form zu kennzeichnen. Art. 35 (1) k DSA) regelt, dass KI-generierte Inhalte, die „Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähnelt“ sichtbar als KI-generiert gekennzeichnet werden müssen. Dennoch fordern Analysten spezifischere Regelungen und Maßnahmen zur effektiven Umsetzung. So gilt etwa diese Bestimmung des DSA nur für sehr große Plattformen und Suchmaschinen[162], nicht jedoch generell.

Verhaltenskodex für Politiker:innen

Nicht im selben Ausmaß verbindlich wie die oben genannten gesetzlichen Maßnahmen wäre Soft-Law. Es gibt bereits den einen oder anderen Verhaltenskodex, etwa der Strengthened Code of Practice on Disinformation auf EU-Ebene[141] oder den Voluntary Code of Practice on Disinformation and Misinformation in Australien (beschrieben in Matasick et al. 2024, ch. 2.2), der die großen Plattformen selbstverpflichtet. Zur „Selbstregulierung der Plattformen“ und deren Wirksamkeit gibt es bereits viele Analysen (siehe beispielsweise in Karaboga et al. 2024, S. 209ff). Im Sinne der Eigenverantwortung der Politik sind solche Abmachungen insbesondere auch zwischen Politiker:innen denkbar. Es ist zumindest ein solches Beispiel bekannt, nämlich der KI-Kodex der Schweizer Mitte-Links-Parteien[139], in dem sich diese dazu gegenseitig verpflichten, auf KI für Negativkampagnen zu verzichten. Solche Verhaltenskodices sind prinzipiell in allen Staaten bzw. auf allen Ebenen, von der EU bis zur Gemeinde denkbar. Darin könnten

---

[161] D. Helbing argumentiert, dass die Abschaffung von Microtargeting *für Parteien* deswegen nicht (allein) zielführend sei, weil „Meinungsbildung ja auch außerhalb der Parteienlandschaft stattfindet“ ***derstandard.at/story/3000000214315/computersoziologe-wir-haben-wahlmanipulation-in-grossem-stil***.

[162] Das sind aktuell nur 20 Plattformen, darunter Google, Meta, TikTok, Twitter/X, siehe ***digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses***.

Regeln über den Gebrauch von Generativer KI im politischen Geschäft, inklusive Wahlkämpfen, vereinbart werden. Regelungsinhalte könnten einerseits die bereits oben genannten *Verbote von Microtargeting, Deepfakes und sonstiger Desinformation* sein. Andererseits könnten *Budgetbeschränkungen* für den Einsatz von Generativer KI im Wahlkampf vereinbart werden, damit wenigstens der Großteil der politischen Kommunikation weiterhin durch Menschen erfolgt (Kruschinski 2023 ab Minute 38). Um solche Verhaltenskodices in der Praxis wirksam werden zu lassen, wurde unter anderem vorgeschlagen, einen *Digitalen Ordnungsruf* bei Missachtung der Regeln ähnlich Ordnungsrufen im Parlamentsplenum durch den/die Parlamentspräsidenten/in einzuführen. Diese niederschwellige Sanktion könnte entweder in der Geschäftsordnung des Parlaments, Landtags oder Gemeinderats festgeschrieben oder bei einer sonstigen unabhängigen Stelle, ähnlich dem Presserat, eingerichtet werden (Karmasin et al. 2024, S. 35).

Internationale Ebene entscheidend

Während manche der o. g. Optionen auch innerhalb Österreichs im Sinne einer verantwortungsvollen Politikgestaltung aktiv und selbstständig umgesetzt werden könnten (insb. Verhaltenskodices), scheint klar, dass die meisten Maßnahmen nur auf internationaler Ebene beschlossen und dort umgesetzt werden können. Dies wird auch von einigen internationalen Akteuren so gesehen. So argumentiert etwa die European Group on Ethics in Science and New Technologies: „More coherent regulation is needed to make digital practices serve people and communities" (EGE 2023, S. 47; siehe auch Matasick et al. 2024). Vor diesem Hintergrund ist eine vielfach genannte Option die (pro-)aktive Beteiligung österreichischer Akteure an der Schaffung internationaler Normen und Regeln in den Bereichen Desinformation und Cyberkriminalität.

6.2 ORGANISATORISCHE UND SONSTIGE ANSÄTZE

Viele Handlungsoptionen, um negative Auswirkungen Generativer KI auf die Demokratie zu begrenzen bzw. positive Aspekte auszuschöpfen, spielen sich abseits oben (6.1) diskutierter Ansätze der Regulierung ab. Neben den im nächsten Abschnitt (6.3) beschriebenen technischen Maßnahmen sind es insbesondere Ansätze, die auf die organisatorische und institutionelle Ebene abzielen, siehe Tabelle 6.

Tabelle 6: Organisatorische und institutionelle Ansätze im Überblick

Stoßrichtung	Maßnahmen
Kompetenzen (Demokratie, Medien, KI) stärken	Bewusstseinsförderung/Prebunking; Überarbeitung der Lehrpläne; Einschlägige Erwachsenenbildung; Einsetzung von Räten unter Einbeziehung von Bürger:innen und Expert:innen; Einrichtung von Demokratielaboren
Informationslandschaft stärken	Förderung der Medienvielfalt und -qualität; Stärkung von Initiativen zur Informationsintegrität; klare Trennung behördlich-offizieller von sonstiger Kommunikation; Gründung nicht-kommerzieller, demokratie-verträglicher Sozialer Medien
Expertise bündeln und Kooperieren	Thematische Vernetzung von Behörden; Stärkung der Aktivitäten zu Foreign Interference; Stärkung internationaler Kooperationen
Stärkung von Forschung und Zivilgesellschaft	Förderung der Zusammenarbeit von Wissenschaft und Zivilgesellschaft; Öffentliche Förderung für unabhängiges Fact-Checking; Monitoring der Informationslandschaft; Förderung der Forschung zu KI und Demokratie

6.2.1 KOMPETENZEN STÄRKEN

Viele Beobachter:innen sind sich einig, dass im Spannungsfeld von Generativer KI und Demokratie die Stärkung verschiedener Fähigkeiten eine zentrale Rolle einnimmt, um die Gesellschaft als Ganzes resilienter zu machen (Matasick et al. 2024). Konkret werden häufig Technologie- bzw. im speziellen KI-Kompetenz, Medienkompetenz und Demokratiekompetenz erwähnt (Bieber et al. 2024; FORWIT 2024; FORWIT/Beirat für Künstliche Intelligenz 2024; Grünke et al. 2024).

Demokratiekompetenz

Demokratiekompetenz definieren Grünke et al. (2024) dabei als „die Fähigkeit der Bürger:innen, aktiv und selbstbestimmt am demokratischen Prozess teilzunehmen, informierte Entscheidungen zu treffen und sich in einem pluralistischen Umfeld zurechtzufinden." (S. 14) Weiters sind die Suche nach Wahrheit und die Wissenschaft als Referenzsystem für Fakten essentielle Bestandteile einer liberalen Demokratie (Grünke et al. 2024).

Medienkompetenz

Besonders die Fähigkeit, unbeeinflusste und informierte Entscheidungen zu treffen, bedarf wiederum einer umfassenden Medienkompetenz, also dem selbstbestimmten und verantwortungsvollen Umgang mit Medien (Grünke et al., 2024). Dazu gehört ganz insbesondere die Fähigkeit zum kritischen Einordnen von Informationen, etwa indem Quellen hinterfragt werden und Informationen mit anderen vertrauenswürdigeren Stellen abgeglichen werden (Bieber et al. 2024). Gerade im Hinblick auf KI-basierte Desinformation ist diese Fertigkeit besonders wichtig. Im Zeitalter von Online-Plattformen, auf denen die Verbreitung von Informationen in Form von Posts, Videos und dergleichen stark algorithmisch vermittelt ist, ist es auch wichtig, diese Verbreitungsmuster zu verstehen – die je nach Plattform sehr spezifisch aussehen können (so sind die Mechanismen auf z. B. TikTok und Facebook sehr verschieden).

Prebunking

Eine speziell auf Falschmeldungen und Desinformation zielende Methode ist das sogenannte Prebunking. Während Debunking (s. u.) auf konkrete, bereits im Umlauf befindliche Falschmeldungen reagiert, geht es beim Prebunking darum, Menschen auf mögliche Falschinformationen vorzubereiten, mit dem Ziel diese zu erkennen und ihre Wirkungsweise verstehen können – noch bevor sie damit in Kontakt kommen. Typische Desinformationstechniken wie Panikmache oder Dekontextualisierung stehen dabei im Mittelpunkt (Upgrade Democracy 2023). Dieses sog. „Impfen gegen Fakenews"[163] wird beispielsweise durch Lehrvideos realisiert, die zunächst unterschiedliche Manipulationstechniken anwenden und anschließend enttarnen (Roozenbeek et al. 2022).

KI-Kompetenz

Generative KI hat besonders im Medienbereich weitreichende Auswirkungen und vielfältige Einsatzbereiche. Daher ist in weiterer Folge ausreichendes Wissen über KI wichtig, um als Einzelne:r die Veränderungen der Medien adäquat zu verstehen und Inhalte richtig einordnen zu können. Dazu gehört einerseits, zu verstehen, wie (Generative) KI funktioniert, was sie (derzeit) kann und was nicht, sie kritisch einordnen zu können, auch durch den kräftig betriebenen Hype rund um das Schlagwort „KI" durchzublicken (zur Vertiefung mit Schwerpunkt Wissensarbeit Strauß/Udrea 2024). Es darf hier aber nicht bei der Aufklärung über die Funktionsweise von KI gestoppt werden. Vielmehr sollten vor allem auch gesellschaftliche Implikationen der Effekte über die persönlich-individuelle

[163] *kas.de/de/web/politische-bildung/politsnack/detail/-/content/prebunking-kann-man-gegen-fake-news-impfen*.

(nicht-)Betroffenheit hinaus in den Blick genommen werden, also etwa in Bezug auf Wirtschaft und demokratische Prozesse. In diese Richtung hat sich Österreich bereits unter dem Schlagwort Digitaler Humanismus[164] profilieren können und die dazu entwickelten Aktivitäten bieten eine gute Basis. Weiters gehört dazu aber auch die Fähigkeit, KI aktiv und verantwortungsbewusst zu nutzen, um die kreativen Potentiale zu erkunden (Grünke et al. 2024). Der (österreichische) Rat für Forschung, Wissenschaft, Innovation und Technologieentwicklung empfiehlt gemeinsam mit dem Beirat für Künstliche Intelligenz der Bundesregierung (2024) dementsprechend auch den Aufbau eines unabhängigen KI-Kompetenzzentrums, das privaten wie öffentlichen Organisationen bei der Weiterbildung sowie beim Experimentieren mit den neuesten Entwicklungen eine gut ausgestattete Infrastruktur und Expertise bietet.

OECD-Empfehlung

An dieser Stelle sei darauf hingewiesen, dass die OECD auch angeregt, bei Entscheidungsträger:innen und Amtsträger:innen auf allen politischen Ebenen die Kapazitäten auszubauen, Desinformation und Misinformation zu erkennen, zu beobachten und deren Verbreitung zu verhindern – welche folglich auch obiger Kompetenzen bedürfen (Matasick et al. 2024).

Lehrpläne in Schulen, Universitäten und in der Erwachsenenbildung

Zur Stärkung dieser Fähigkeiten gibt es eine Vielfalt an Optionen, die sowohl von der öffentlichen Hand als auch durch private Initiativen gesetzt werden können. Naheliegenderweise ist eine häufige Empfehlung, Lehrpläne zu ergänzen sowie – um diese auch effektiv umzusetzen – Lehrpersonal entsprechend weiterzubilden (Matasick et al. 2024). Eine weitere mögliche Maßnahme ist, ein breiteres Bildungsangebot auch für Erwachsene aufzubauen (Bieber et al. 2024). Als positives Beispiel kann hier das Erweiterungscurriculum Digitale Kompetenzen an der TU Wien angesehen werden, wo Studierende, die nicht Informatik studieren, sich umfangreiche Kompetenzen aneignen können.[165]

Bereits erste Initiativen und Projekte

Das ITA-Projekt „Critical Artificial Intelligence Literacy“ (CAIL) hat mit seinem CAIL-Framework einen Beitrag zur Vermittlung von KI-Wissen in Betrieben geleistet (Strauß/Udrea 2024). Das laufende internationale Projekt „Artificial Intelligence and the Shaping of Democracy“ (A.I.D.) mit ITA-Beteiligung widmet sich der Entwicklung von Bildungskonzepten im Umgang mit KI-Technologien.[166] Auch das laufende österreichische Leitprojekt im Bereich KI „Fostering Austria’s Innovative Strength and Research Excellence in Artificial Intelligence“ (FAIR-AI) erarbeitet u. a. Lehrplan-Module für die unterschiedlichen Bildungsstufen.[167] Weiters sind Leitfäden für Bürger:innen (sog. Citizen Guides, siehe Villar García et al. 2021, S. 106f) zur Erkennung von Desinformation zu empfehlen. Ein Beispiel in vielen Sprachen ist die Information des European Fact-Checking Standards Networks (EFCSN, mehr dazu weiter unten) AI@Elections,[168] das in 27 Sprachen verfügbar ist.

Breit zusammengesetzte Räte

Darüber hinaus ist eine wiederkehrende Empfehlung die Schaffung von Räten, die unter Einbeziehung von Lai:innen, Expert:innen und Stakeholdern Lösungsvorschläge erarbeiten oder Prozesse begleiten und „durch das bestmögliche verfügbare Fachwissen unterstützt werden“ sollten (Europäische Kommission 2023, S. 13). Wichtig bei derartigen Räten ist, dass auch die Perspektiven besonders schutzbedürftiger und marginalisierter Personengruppen ausreichend Be-

164 *digitalhumanism.at*.
165 *informatics.tuwien.ac.at/bachelor/digitale-kompetenzen/*.
166 *oeaw.ac.at/ita/projekte/aid*.
167 *oeaw.ac.at/ita/projekte/fair-ai*.
168 *efcsn.com/projects/ai_euelections/*.

achtung findet. So präsentiert etwa ‚upgrade democracy' die Idee eines Rats für digitale demokratische Diskurse (Gabriel/Hadeed 2024). Diesem Rat gehörten neben Wissenschaftler:innen und per Los ausgewählte repräsentative zusammengesetzte Bürger:innen auch Vertreter:innen von Industrieverbänden, Medien und Verbraucherschützer:innen an. Ziel dieses Rats wäre zu erörtern, „welche Optionen und Standards die rechtliche und algorithmische Moderation von Diskursen im demokratischen Kontext umfassen sollte." (Gabriel/Hadeed 2024, S. 36) Medienunternehmen und Onlineplattformen könnten sich in der Folge entscheiden, diesen Standards zu folgen oder sich mit Begründung dagegen zu entscheiden – was in weiterer Folge zur weiteren Debatte beitragen kann. Durch dieses Modell würden alle drei oben angeführten Kompetenzfelder gestärkt. Ein ähnliches Konzept schlägt Coeckelbergh (2024) vor. In seiner Vision berät ein ständiger KI-Rat unter Einbeziehung von Lai:innen und Expert:innen die Regierung und spricht Empfehlungen aus. Auch hier geht es neben der Erarbeitung von gut informierten Vorschlägen um eine Stärkung der Partizipation und der Demokratie.

Wissenschafts- und Demokratielabore

In den kürzlich veröffentlichten Empfehlungen für die XXVIII. Legislaturperiode schlägt FORWIT (2024) die Einrichtung von Wissenschafts- und Demokratielabors in jedem Bundesland vor. Diese Labore stünden allen Bürger:innen offen, wären breit interdisziplinär aufgestellt und würden der Wissenschafts- und Demokratievermittlung dienen. Als Ausgangsbasis dafür könnte die Initiative DNAustria[169] des Bundesministeriums für Bildung, Wissenschaft und Forschung dienen. In diesem Sinne ist auch der kürzlich vorgestellte Plan, in der Aula der Wissenschaften ein Kompetenzzentrum für Wissenschaftsvermittlung aufzubauen,[170] ausdrücklich zu begrüßen. Auch die Demokratiewerkstatt mit ihren Medienwerkstätten im österreichischen Parlament ist als bestehendes Angebot hervorzuheben und könnte als Vorbild für weitere Angebote dieser Art dienen.

Stärkung der Forschung zu KI und Demokratie

In alle obigen Vorschläge fließt fundiertes, auf wissenschaftlichen Prinzipien aufbauendes Wissen ein. Um diesen Wissensbestand zu erhalten und zu erweitern, also um Problemstellungen rund um Generative KI und Demokratie tiefgreifend zu bearbeiten, wird die Stärkung der Forschung in diesem Bereich genannt. Dies gilt besonders für die Frage, wer weshalb für Desinformation besonders anfällig ist, um die Prävention entsprechend zu gestalten (Matasick et al. 2024).

6.2.2 INFORMATIONSLANDSCHAFT STÄRKEN

Die Stärkung der Informationslandschaft (engl. information ecosystem) ist eine weitere wichtige Stoßrichtung, um das demokratische Gefüge gegenüber den Gefahren der Generativen KI resilienter zu machen.

Informationsintegrität

Ein zentraler Aspekt ist es, die Informationsintegrität zu bewahren bzw. auszubauen, d. h. die Verfügbarkeit von frei zugänglichen, verlässlichen, evidenzbasierten und pluralistischen Informationsquellen, um so eine informierte Entscheidung der Bürger:innen zu gewährleisten (Matasick et al. 2024). Die OECD empfiehlt explizit, nicht Inhalte staatlich zu regulieren (denn dies könnte mit demokratischen Grundrechten in Konflikt geraten), sondern stattdessen die Bedingungen der Medienlandschaft so zu gestalten, dass die genannten Ziele leichter erreichbar sind (Matasick et al. 2024) und freien Journalismus als wichtigen

Pluralistische Medienförderung mit Fokus Integrität

[169] *dnaustria.at*.

[170] *oeaw.ac.at/news/oesterreichs-groesstes-science-communication-center-entsteht-in-wiener-innenstadt-1*.

Player gegen Desinformation zu unterstützen (Villar García et al. 2021, S. 122f). Daher ist eine Medienförderung mit einem Fokus auf Quellenpluralität, Verantwortlichkeit und Transparenz ein möglicher Ansatz zur Stärkung der Informationsintegrität (Karmasin et al. 2024, S. 35f).

Trennung offizieller von sonstiger Kommunikation

Eng damit verknüpft ist auch der Hinweis der OECD, öffentliche Informationskanäle robust aufzustellen, d. h. offizielle Kommunikation klar von parteipolitischer und von Partikularinteressen geleiteter Kommunikation abzugrenzen (Matasick et al. 2024). Dadurch wird das Vertrauen der Bevölkerung in die Verlässlichkeit staatlicher Kommunikation gestärkt, da der Anschein von Interessenskonflikten zwischen notwendiger Information der Bevölkerung und politisierten Interessen minimiert wird.

Faktenchecks

Der wohl wichtigste Beitrag zur Stärkung der Informationsökologie ist das Faktenchecken. Damit ist gemeint, dass im Internet kursierende Meldungen daraufhin untersucht werden, ob sie den Tatsachen entsprechen oder es sich um Mis- oder Desinformation handelt. Dabei wird durch genaue Analyse der Quellen, Quervergleiche, das Nachvollziehen der Verbreitungswege, Bild- und Textvergleiche, zeitliche Analysen usw. versucht, den Wahrheitsgehalt zu erkennen. Faktenchecks enden in der Regel mit der Veröffentlichung von kurzgefassten Analysen im Internet, i. d. R. auf denselben Kanälen, auf denen die ursprüngliche Meldung verbreitet wurde. Die Austria Presse Agentur sammelt entsprechende Faktenchecks, inklusive Erklärung, wie der Check durchgeführt wurde.[171] Die APA ist auch das österreichische Mitglied in „Elections24Check", einer Initiative des Zusammenschlusses von 40 europäischen Factchecking-Organisationen EFCSN[174], die sich auf die Wahlen im Jahre 2024 konzentriert haben.[172] Ein Nachteil von Factchecking ist leider, dass es ein relativ langsames und reaktives Verfahren ist – d. h. teils werden Falschinformationen weit verbreitet, bevor ein Faktencheck veröffentlicht werden kann. Darüber hinaus ist Factchecking für die Nutzer:innen insofern unpraktikabel, da es relativ zeitaufwändig ist und wenige im Alltag die Ressourcen haben, Meldungen ständig auf ihren Wahrheitsgehalt hin zu überprüfen. Daher ist es besonders wichtig, hierfür in erster Linie Medien und in weiterer Folge auch NGOs ausreichend zu unterstützen, um diese Bürde nicht den einzelnen aufzuladen sowie unterstützende Werkzeuge zu entwickeln (dazu weiter unten mehr).

Trust-Checking

Als komplementär zu Factchecking gibt es auch die Praxis des sogenannten Trust-Checkings: Hier wird Information nicht direkt inhaltlich geprüft, sondern hinsichtlich bestimmter Metakriterien untersucht, also etwa „Qualitätskriterien wie dem Vorhandensein von Quellenangaben, korrekten Zitaten oder authentischen Bildern auf ihre Glaubwürdigkeit hin überprüft. Erfüllt eine Information diese Kriterien nicht, wird sie als weniger glaubwürdig eingestuft." (Bertelsmann Stiftung 2023). Siehe dazu auch die Idee des „trusted flagging" weiter unten.

Debunking

Etwas anders als Factchecking ist das sogenannte Debunking. Hier konzentriert man sich gewissermaßen auf einer abstrakteren Ebene auf Narrative statt einzelne Fakten. Ziel ist es hier, Desinformationsmuster und Verschwörungserzählungen zu größeren Themen wie Gender oder Klima aufzuzeigen, indem Quellen und Inhalte eingeordnet werden sowie glaubwürdige Alternativquellen hervorgehoben werden (Upgrade Democracy 2023).

---

171 *apa.at/service/faktencheck-2/*.

172 *elections24.efcsn.com/*.

Stärkung der Zusammenarbeit von Wissenschaft, Zivilgesellschaft, Medien und Onlineplattformen

Durch die Stärkung der Zusammenarbeit von Wissenschaft, Zivilgesellschaft, Medien und Onlineplattformen kann Desinformation, besonders durch verschiedene Fälschungen, am effektivsten entgegengewirkt werden. Ein Beispiel ist das German-Austrian Digital Media Observatory (GADMO),[173] der größte Zusammenschluss von Faktencheck-Organisationen und Forschungsteams gegen Desinformation in Deutschland und Österreich, bestehend aus den Nachrichtenagenturen APA, DPA, AFP und der Faktencheckorganisation Correctiv mit Unterstützung von Technologie-Spezialisten, wie u. a. dem Austrian Institute of Technology (AIT). GADMO ist die zentrale Plattform für deutschsprachige Faktenchecks, hier werden unter anderem Desinformationskampagnen, deren Verbreitung und Gegenmaßnahmen großer Technologiekonzerne beobachtet und Medienkompetenz in Deutschland und Österreich vermittelt. GADMO ist Teil eines EU-weiten Netzwerks aus 14 Hubs, die Faktenchecks in der EU sicherstellen. Beim European Fact-Checking Standards Network (EFCSN)[174] findet sich eine Liste weiterer Organisationen, die den strengen Prüfungskriterien für Faktenchecker:innen entsprechen. Im Vorfeld der EU-Parlamentswahlen 2024 wurde eine dezidierte Fact-Checking-Webseite vom EFCSN in Betrieb genommen.[175]

Öffentliche Unterstützung für Faktenchecks?

Es ist vorstellbar, Faktenchecks nicht den Akteuren (z. B. Presseagenturen, aber auch privaten NGOs.) alleine zu überlassen, sondern auch staatlich zu unterstützen, etwa durch Subventionen, durch Bereitstellung von Infrastruktur oder durch einen öffentlich-rechtlichen Rahmen, ähnlich dem öffentlich-rechtlichen Rundfunk, wie er in vielen Staaten, so auch in Österreich, etabliert ist. Die Grenzen privater Initiativen zeigen sich etwa deutlich an *mimikama – Verein zur Aufklärung über Internetmissbrauch,* einem bekannten zivilgesellschaftlichen Akteur in Österreich. Mimikama hat sich der Aufklärung von Internet-Nutzer:innen über Internetkriminalität und Internetbetrug verschrieben, in Form von Veröffentlichungen auf der eigenen Webseite.[176] Weiters kooperiert Mimikama mit Medien und Bildungseinrichtungen und betreibt eine Beschwerdestelle für illegale Internet-Inhalte. Der Verein sieht sich u. a. als Plattform für Erfahrungsaustausch. Nicht-Regierungsorganisationen kommt auch gerade in Bezug auf Factchecking eine besondere Rolle zu, da ihnen oft, gerade von marginalisierten Bevölkerungsgruppen, mehr vertraut wird als offiziellen Behörden (Matasick et al. 2024). Jedoch stellt sich immer wieder deutlich heraus, dass die NGO-Schiene sehr prekär ist, da die Finanzierung aufwändiger Recherchen und damit von Personal nur durch Spenden schwer aufzubringen ist. Dabei kommt im Digital Services Act von Online-Plattformen unabhängigen Organisationen in Form von vertrauenswürdigen Hinweisgebern, sog. *Trusted Flagger*,[177] eine spezielle Rolle zu: Hinweise solcher registrierter vertrauensvoller User:innen müssen von Onlineplattformen vorrangig behandelt und schneller behandelt werden als von „üblichen" Nutzer:innen, da sie aufgrund ihrer Expertise vertrauenswürdiger einzuschätzen sind.

Trusted Flagger

Nicht-kommerzielle, demokratieverträgliche Soziale Medien

Ein noch weitergehender Vorschlag lautet, in Europa nicht kommerzielle, eventuell öffentlich-rechtliche Soziale Medien zu gründen, die sich – anders als die von den USA oder China dominierten Plattformen strikt an den hier geltenden Normen halten und deren Algorithmen sich nicht an Profitmaximierung und

[173] *gadmo.eu*.
[174] *efcsn.com*.
[175] *elections24.efcsn.com*.
[176] *mimikama.org*.
[177] *digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa*.

daher Verstärkung von aufwiegelnden, hassverstärkenden und eventuell nicht faktenbasierten Posts orientieren. So ein neues Medium könnte einen möglichst rationalen, demokratieverträglichen Diskursraum in Europa schaffen. Dirk Helbing, Professor für Computational Social Science, formuliert dies etwa so:

> *„Wie viele Milliarden hat man investiert in Cybersecurity? Und wie viele in ein digitales Upgrade der Demokratie? Da gibt es ein großes Ungleichgewicht. […] Wir bräuchten nicht-kommerzielle Demokratieplattformen, aber stattdessen haben Tech-Firmen wie Facebook/Meta, die Räume für Meinungsbildung besetzt, die eigentlich öffentlich hätten sein müssen. So werden unsere Grundrechte zu Geld gemacht."* [178]

Christian Ehler, Europaabgeordneter der EVP, und Matthias Pfeffer, Founding Director beim Council for European Public Space schreiben:

> *„Desinformationskampagnen von innen und außen haben zuletzt den Europawahlkampf geprägt. Sie zeigen erneut, wie wichtig es ist, dass Europa endlich einen resilienten öffentlichen Raum im Digitalen schafft. Dezentrale Plattformen, auf denen seriöse Nachrichten aus Europa in alle europäischen Sprachen übersetzt und allen Bürgern zugänglich gemacht werden, sind heute technisch möglich. Die Politik muss nur wollen."* [179]

Die Finanzierung könnte ähnlich wie im Rüstungssektor Airbus über ein Public-private-Partnership oder wie im Medienbereich rein öffentlich-rechtlich wie der TV-Sender ARTE organisiert sein.[180] Ähnliches forderte auch das *Public Service Media Manifesto* (2021/2022).[181] Der Erfolg einer solchen Initiative hinge freilich von der Attraktivität ab, die entscheidend wäre, um Bürger:innen in wirklich großen Zahlen zum Wechsel von den etablierten Plattformen wie z. B. „X" (vormals Twitter) zu animieren. Der im Herbst 2024 stattgefundene Wechsel vieler prominenter Twitter/X-Nutzer:innen zu Bluesky[182] zeigt, dass beim Vorhandensein guter Alternativen ein Umstieg auch auf ein Public-Service-Medium durchaus realistisch wäre.

Monitoring der Informationslandschaft

Eine weitere Stoßrichtung, um die Informationslandschaft resilienter zu gestalten, ist das Monitoring der Informationslandschaft durch unabhängige Initiativen. Derzeit arbeitet etwa das Projekt *Upgrade Democracy*[183] der Bertelsmann Stiftung an Lösungen, die sowohl Desinformation in ihren jeweiligen Kontexten effektiv bekämpfen als auch innovative digitale Werkzeuge nutzen, um die Demokratie zu fördern. Durch das Monitoring digitaler Plattformen können (unabhängige) Forschende und zivilgesellschaftliche Organisationen den Ursprung, die Verbreitung, den Kontext, die beteiligten Akteure, die Wirkung und weitere Aspekte von Desinformationskampagnen beobachten und analysieren (Bertelsmann Stiftung 2023). Ein Beispiel für ein solches Werkzeug zur Überwachung von Nachrichtenplattformen ist die Website *Debunk.org*.[184]

---

178 Der Standard vom 7.4.2024, ***derstandard.at/story/3000000214315/computersoziologe-wir-haben-wahlmanipulation-in-grossem-stil***.

179 Tagespiegel, 8.8.2024, ***background.tagesspiegel.de/digitalisierung-und-ki/briefing/daseinsvorsorge-fuer-die-demokratie***;
siehe auch A. Thurnher am 9.3.2024, ***falter.at/seuchenkolumne/20240309/claudia-plakolm-wird-digitalstaatssekretaerin-helau***.

180 So bspw. A. Thurnher, 11.11.2024: ***falter.at/seuchenkolumne/20241111/donaldology-iii-ich-muss-die-muskfrage-stellen***.

181 ***ia902206.us.archive.org/5/items/psmi_20220127/psmi.pdf***.

182 ***orf.at/stories/3376168/***.

183 ***bertelsmann-stiftung.de/de/unsere-projekte/upgrade-democracy***.

184 ***debunk.org***.

KI-basiertes Entkräften von Falschinformationen

Das Entkräften von Falschinformationen kann sehr zeitaufwendig sein, wenn es von Einzelpersonen durchgeführt wird, die Nachrichtenplattformen einzeln durchsuchen. Durch den Einsatz Generativer KI, wie es Debunk.org, eine nichtstaatliche und unabhängige Organisation, tut, wird der Prozess beschleunigt und hilft, einen großen Teil der Falschinformationen aufzudecken (Debunk 2023). In dieser Art Automatisierung besteht prinzipiell großes Potenzial für die Zukunft.

Kennzeichnung KI-generierter Inhalte

Abschließend sind alle Initiativen, die eine umfassendere Transparenz durch Kennzeichnung von KI-generierten Inhalten bzw. auch zum Nachweis von Inhalten, die definitiv nicht von KI generiert wurden (hier v. a. im Falle von Fotografien, Audio- und Videoaufnahmen) begrüßenswert. Detailliert wurden die Möglichkeiten und bestehende Bemühungen, aber auch deren etwaige Nachteile bereits in Abschnitt 2.3 besprochen, siehe auch 6.1.

6.2.3 KOOPERATIONEN UND BÜNDELUNG VON EXPERTISE

Da Generative KI und Demokratie auf komplexe Weise zusammenhängen, sind die Zuständigkeiten und Expertisen über verschiedene Sektoren und Behörden verteilt. Beobachter:innen sind sich daher einig, dass es notwendig ist, relevante Expertise zu bündeln bzw. effektiv zu vernetzen (z.B. FORWIT 2024). Da Generative KI besonders online und damit global eingesetzt und verbreitet wird, sind internationale Kooperationen – und hier speziell unter (liberalen) Demokratien – essentiell. Durch Partnerschaften und Zusammenarbeit kann einerseits koordiniert gegen Desinformationskampagnen vorgegangen werden. Andererseits dienen diese aber auch dem Austausch über Methoden und Strategien, um Best Practices zu entwickeln und demokratische Strukturen und Prozesse zu stärken (Matasick et al. 2024).

Aktivitäten auf EU-Ebene

Österreich und die Europäische Union sind in dieser Hinsicht bereits relativ gut aufgestellt. Besonders auf EU-Ebene existieren einige Aktivitäten besonders zu den Themen Desinformation und Hybride Bedrohungen. So wird seit 2017 die EU Cyber Diplomacy Toolbox[185] entwickelt, die in Österreich im Rahmen der nationalen Cyberstrategie[186] ihren Anschlusspunkt hat. Seit 2022 wird des Weiteren an der EU Hybrid Toolbox gearbeitet,[187] die koordinierte Antworten der EU auf hybride Kampagnen unterstützen soll. Erst 2024 hat der Europäische Rat einen Orientierungsrahmen für die Einrichtung sogenannter Hybrid Rapid Response Teams beschlossen. Diese Teams sollen „auf Ersuchen zur Vorsorge gegen hybride Bedrohungen und Kampagnen sowie zu deren Abwehr eingesetzt werden können“ und somit als Service für Mitgliedsländer sowie Partnerländer dienen).[188]

Das Hybrid Centre of Excellence for Countering Hybrid Threats ist eine seit 2017 operative Stelle zur Koordinierung der Aktivitäten von 36 Mitgliedsstaaten und damit über die EU hinaus.[189] Kernaufgabe ist, die Demokratien der Mitglieds-

---

185 *cyber-diplomacy-toolbox.com*.

186 *onlinesicherheit.gv.at/Services/Publikationen/Sicherheitsstrategien-und-Initiativen/2021-Oesterreichische-Strategie-Cybersicherheit.html*.

187 *consilium.europa.eu/en/press/press-releases/2022/06/21/council-conclusions-on-a-framework-for-a-coordinated-eu-response-to-hybrid-campaigns/*.

188 *consilium.europa.eu/de/press/press-releases/2024/05/21/hybrid-threats-council-paves-the-way-for-deploying-hybrid-rapid-response-teams/*.

189 Siehe *hybridcoe.fi/about-us/*.

staaten vor äußeren Einflüssen im Sinne hybrider Bedrohungen zu schützen, u. a. indem deren Fähigkeiten, diese zu verhindern und zu bekämpfen, gestärkt werden.

Eine gute Koordination könnte Antworten erlauben, die über die Fähigkeiten einzelner Staaten hinausgehen (Lasoen 2022). Bereits seit 2019 bietet das Rapid Alert System[190] eine erste Grundlage für Informationsaustausch und Koordination. Eine erfolgreiche Koordination ist kein Selbstläufer: *„Die Umsetzung wird ein echter Test für die Fähigkeit der EU sein, über die Grenzen zwischen den Bereichen der inneren und äußeren Sicherheit sowie über verschiedene Politikbereiche hinweg zu handeln."* (Lasoen 2022, S. 10). Auch die oben angesprochenen Rapid Response Teams sind auf (zivile wie militärische) Expertise der Mitgliedsländer angewiesen.[188]

Strategic Communication

Unter dem Stichwort Strategic Communication wird darüber hinaus ein weiterer wichtiger Ansatz gegen hybride Bedrohungen diskutiert. Dabei handelt es sich um eine strategische, d. h. längerfristige Kommunikationsstrategie mit klaren Zielen, die sich nicht an eine allgemeine Öffentlichkeit richtet und auch nicht reaktiv (wenngleich flexibel-adaptiv), sondern pro-aktiv ist und sich an multiplen und diversen Zielgruppen ausrichtet (Villar García et al. 2021). Langfristig heißt hier vor allem, dass kontinuierlich, geplant und über einen längeren Zeitraum Aktivitäten gesetzt werden – und über diesen Zeitraum die Kommunikation (und Aktionen als Teil der Kommunikation) konsistent gehalten wird. Damit wird über reaktives Verhalten auf Vorfälle hinaus gegangen. Essentiell ist aber auch, dass die Informationen vertrauenswürdig, glaubhaft und faktenbasiert sind – und schließlich auch demokratischen Werten folgen müssen. Auch dieser Ansatz erfordert einen hohen Koordinierungsaufwand der beteiligten Akteure.

Bisherige Aktivitäten in Österreich

Daraus lässt sich schließen, dass auf EU-Ebene bereits einige Anschlusspunkte bestehen, die aber den Aufbau relevanter Expertise in Österreich nicht ersetzen, sondern nur ergänzen können. In dieser Richtung gibt es bereits die gesamtstaatliche, interministerielle AG Hybrid, die auch eine eigene „Kerngruppe Desinformation" umfasst.[191] Diese Kerngruppe war besonders im Vorfeld der EU-Wahlen 2024 verstärkt aktiv.[192] Derartige Bündelungen von Kompetenzen und Expertise im Austausch mit Partnerländern beizubehalten und mit ausreichend Ressourcen für wirkungsvolle Arbeit auszustatten erscheint eine naheliegende und wichtige Handlungsoption.

Forschung und Entwicklung

Zu diesen Ressourcen zählen auch entsprechende technische Entwicklungen und Forschungen. Ein Beispiel in diese Richtung ist das Projekt EUvsDisinfo[193] des Europäischen Auswärtigen Dienstes, das seit 2015 kremlnahe Desinformationskampagnen beobachtet, analysiert und Gegenstrategien entwickelt. Dieses Projekt hat auch zum oben erwähnten Rapid Alert System geführt. Aber auch in Österreich entwickelte Werkzeuge und Strategien, etwa im Rahmen des Sicherheitsforschungsprogramms KIRAS[191] leisten einen wichtigen Beitrag.

---

190 *eeas.europa.eu/node/59644_en*.

191 17721/AB – Anfragebeantwortung BM Tanner, 12.6.2024, *parlament.gv.at/gegenstand/XXVII/AB/17721*.

192 18342/AB – Anfragebeantwortung BK Nehammer, 21.8.2024, *parlament.gv.at/gegenstand/XXVII/AB/18342*.

193 *euvsdisinfo.eu*.

6.2.4 DEMOKRATIEFÖRDERNDE INNOVATION FORCIEREN UND SOUVERÄNITÄT STEIGERN

Vorreiterrolle der EU

Die Europäische Union hat sich spätestens seit der Datenschutz-Grundverordnung als globaler Vorreiter in Fragen Regulierung von Problemfeldern im Zusammenhang (nicht nur) mit Informationstechnologien positionieren können. Eine ähnliche Vorreiterrolle könnte die EU, und damit Österreich, auch in Bezug auf Innovationen spielen, die demokratische Grundwerte ins Zentrum rücken. Das ist gerade in Zeiten, wo große Innovationsschübe einerseits vom Silicon Valley, andererseits von China kommen und liberale Demokratien weltweit unter Druck stehen, wichtig.

Europäische Plattformen

Bereits im Abschnitt 6.2.2 wurde der Vorschlag, Soziale Medien und Online-Plattformen nach eigener (zu entwickelnder) Logik in Einklang mit liberal-demokratischen, europäischen Grundwerten zu gründen, präsentiert. Denn auch wenn rezente EU-Regelwerke auch nicht-europäischen Plattformen Zugeständnisse abringen, so ergibt sich daraus nicht automatisch ein grundsätzlich anderes Angebot, das dem Kommunikationsbedürfnis der österreichischen und europäischen Bevölkerung gerecht wird. Analog verhält es sich mit anderen technologischen Anwendungsfeldern.

Unterstützung demokratieförderlicher Innovationen

Dementsprechend ist eine gezielte, strategisch geplante und europäisch koordinierte Förderung von Innovationen, die demokratieförderlich sind, eine wichtige Option. Als Inspiration kann die Abteilung „Forschung für technologische Souveränität und Innovation" des deutschen Bundesministeriums für Bildung und Forschung dienen.[194] Um der in den Abschnitten 4.4 und 4.7 konstatierten Machtkonzentration mit Bedrohung der digitalen Souveränität entgegenzuwirken, braucht es eine gezielte Wirtschaftsförderung, um nationale bzw. europäische „Champions" aufzubauen, die ein echtes Gegengewicht gegen Google, Meta, Amazon & Co. bilden könnten, ohne deren Marktpraktiken zu kopieren.

Forschung und Innovationen zu KI und Demokratie

Im Bereich der KI könnte es folglich zur gezielten Förderung von Forschung zu KI und Demokratie kommen. Darauf basierende KI-Innovationen könnten demokratische Werte und Prozesse schützen, auf interdisziplinärer Analyse basierend auch zu demokratisch stärker legitimierten Lösungen führen. Bereits jetzt gibt es erste Ansätze in diese Richtung. Einerseits sei hier die von Wien ausgehende Strömung des Digitalen Humanismus[195] erwähnt, die humanistische Werte ins Zentrum rückt. Andererseits versucht das Cluster of Excellence zu Bilateraler Künstlicher Intelligenz[196] fundamentale Nachteile Generativer KI durch technologische Innovationen zu überwinden – und ihr Fähigkeiten für „echtes" logisches Schließen einzuprogrammieren, was zu vertrauenswürdigerer und verlässlicher KI führen soll.

Open-Public-Data-Strategie

Open Source

Anreize und Infrastruktur für das Sammeln und Bereitstellen von qualitativ hochwertigen und pluralistischen Daten im Rahmen einer Open-Public-Data-Strategie könnte KI-Forschung europäischer Prägung Auftrieb geben – sind doch verlässliche Daten guter Qualität essentiell für gute Resultate. Gleichsam muss sichergestellt werden, dass davon nicht-europäische Akteure nicht überpropor-

194 *bundesregierung.de/breg-de/aktuelles/deepfakes-2246064*.
195 *digitalhumanism.at*.
196 *vcla.at/2024/05/bilateral-artificial-intelligence/* und *fwf.ac.at/aktuelles/detail/oesterreichs-naechste-exzellenzcluster-starten*.

tional profitieren, ohne eine Gegenleistung liefern zu müssen, und dass Datenarbeit nicht in ausbeuterischer Weise in den Globalen Süden ausgelagert wird. Ähnlich verhält es sich mit der Open Source auf der Software-Seite: Auch das kann dazu beitragen, dass die Abhängigkeit von Big Tech reduziert wird.

Langfristiges Engagement und großes Budget notwendig

All diese Ansätze würden auch langfristig zu einem Wissens-, Kompetenz- und Infrastruktur-Aufbau in Österreich und der EU führen. Die größten Nachteile sind aber klar in der zeitlichen Dimension zu sehen: Diese Ansätze brauchen langfristiges Engagement und umfangreiche budgetäre Ressourcen, denn ihre Resultate sind bestenfalls mittel- bis langfristig spürbar. Das ist gerade im schnelllebigen Umfeld der KI ein Risiko – ein Umfeld, das außerdem viel Risikokapital für teils auch fragwürdige Geschäftsmodelle zu mobilisieren vermag.

6.3 TECHNISCHE ANSÄTZE

In diesem Abschnitt werden die Herausforderungen der nur über technische Hilfsmittel ersichtlichen Kenntlichmachung und der automatisierten Detektion u. a. aus ethischer Perspektive aufgezeigt und weitere technische Methoden vorgestellt, die Medienschaffenden und Endverbraucher:innen unmittelbar helfen können, generierte oder manipulierte Inhalte zu erkennen. Damit können diese Methoden auch unterstützend zur Erhöhung der Medienkompetenz beitragen. Darüber hinaus werden weitere technische Ansätze vorgestellt, die manche Risiken von Desinformation oder allgemein Generativer KI adressieren und abmildern können.

Tabelle 7: Überblick Technische Ansätze

Ansatz	Zur Lösung wovon?	Kurzbeschreibung
[Keine Technik]	Feststellen, ob ein Inhalt KI-generiert ist	Abgehen von der Idee, per Technik ausreichend robust Generative KI erkennen zu können bzw. ein Wettrüsten vermeiden
Inhalte mit KI automatisch analysieren		Spezielle KI-Methoden entwickeln, die Artefakte von Generativer KI erkennen können (siehe auch Abschnitt 2.3.2)
Wasserzeichen und Fingerprints		Bei Erstellung von Inhalten diese eindeutig als KI-generiert oder eben nicht KI-generiert kennzeichnen (siehe auch Abschnitt 2.3.1)
Software zur Unterstützung von Fact-Checking	Feststellen, ob etwas Falschinformation ist	Werkzeuge zur Auswertung von z. B. frei zugänglichen Informationen erleichtern Menschen das Fact-Checking
Verbreitungsmuster von Falschinformation erkennen		Durch das Nachvollziehen davon, wie Information im Netz verbreitet wird, können Rückschlüsse über ihre Verlässlichkeit gezogen werden
Benachteiligung von Quellen mit Desinformation oder generierten Inhalten	Verbreitung von Falschinformation eindämmen	Quellen wie Webseiten und Social-Media-Accounts
Slow AI	Gesellschaftliche Risiken von KI minimieren	Generative KI ist zwar derzeit eine sehr populäre, aber letztlich nur eine Spielart von KI – Slow AI stellt die soziale Verantwortung in den Mittelpunkt

6.3.1 KI-GENERIERTE INHALTE ERKENNEN

Eine große Herausforderung in Bezug auf Generative KI ist, feststellen zu können, ob Inhalte mit generativer KI erstellt oder manipuliert wurden. Für eine detailliertere Besprechung des technischen Hintergrunds verweisen wir auf Abschnitt 2.3. In diesem Kapitel fassen wir die Ansätze grob zusammen und beleuchten gesellschaftliche und ethische Implikationen dieser Technologien.

Technische vs. nicht-maschinelle Erkennung von KI-generierten Inhalten

Doch bevor wir einen Blick auf die möglichen technischen Ansätze werfen, sei an dieser Stelle darauf hingewiesen, dass einige Expert:innen dafür plädieren, nicht-maschineller Erkennung von KI-generierten oder -manipulierten Inhalten den Vorzug zu geben. Vertreter:innen dieser Position sehen bei bestehenden technischen Lösungen zu große Unsicherheiten in der Erkennungsleistung. So kann z. B. die Verlässlichkeit von Detektoren nicht gesichert evaluiert werden – man kann nicht unzweifelhaft feststellen, wann ein Erkennungstool an seine Grenzen stößt oder wo diese liegen. Selbst bessere Erkennungsleistung bietet keinen sinnvollen Schutz, weil sich das Ganze zu einem Wettlauf zwischen Erkennungsdiensten und Produzent:innen von Inhalten entwickeln würde, vergleichbar mit Anti-Viren-Software. Erzeuger:innen von manipulativen Inhalten sind dabei im Vorteil, da sie ihre Inhalte so lange gegen die Erkennungsprogramme optimieren können, bis diese sie nicht mehr als generiert bzw. manipuliert markieren. Dann werden sie veröffentlicht, und erst danach hat die „Abwehr" die Möglichkeit, die Erkennungsleistung an Hand der neu veröffentlichten Materialien zu verbessern. Stattdessen sind Informationen wie der Kontext, in dem Inhalte veröffentlicht und verbreitet werden möglicherweise aufschlussreicher. Das ist insbesondere auch hinsichtlich Einordnung als Desinformation, Misinformation oder gar Satire wichtig:

> *„Die Verwendung von Erkennungsmethoden zur Identifizierung von Deepfakes ist von entscheidender Bedeutung, aber noch wichtiger ist es, die wahre Absicht der Personen zu verstehen, die Deepfakes veröffentlichen. Dies erfordert die Beurteilung der Benutzer:innen auf der Grundlage des sozialen Kontexts, in dem der Deepfake entdeckt wird, z. B. wer ihn verbreitet hat und was die Person dazu gesagt hat."*
> (Nguyen et al. 2022, S. 13; unsere Übers.)

Die Hoffnung ist daher, dass man auf andere, eben nicht-technische Weise manipulierte bzw. generierte Inhalte erkennen kann, und damit auch keine Notwendigkeit bestünde, die Technik so schnell weiterzuentwickeln, wodurch vielleicht auch simple Erkennungsmaßnahmen länger zielführend bleiben. Das bedeutet, dass es sich also um eine Aufgabe für die Bildung handelt (siehe oben 6.2.1).

Zur Authentifizierung und Verifizierung von Inhalten sowie für die Offenlegung und Nachverfolgung des Einsatzes von Generativer KI kommt, den genannten Einwänden zum Trotz, im Allgemeinen technischen Ansätzen viel Aufmerksamkeit zu. Sie sind einerseits eng mit den Regulierungsmaßnahmen für Anbieter und Betreiber dieser Technologien, aber auch mit den Verpflichtungen von Plattformen verknüpft und sollen dazu beitragen, dass Endverbraucher:innen per KI generierte oder manipulierte Informationen von vertrauenswürdigen Inhalten unterscheiden kann.

Zwei Methoden zur Erkennung:

Unter den technischen Ansätzen zur Erkennung von KI-generierten Inhalten können wir grob zwischen zwei Methoden unterscheiden: Erstens kann man versuchen, mit Analysemethoden Artefakte zu identifizieren, die auf Generative KI als Ursprung hindeuten, wie im Abschnitt 2.3.2 detailliert vorgestellt. Je nach Medium (Text, Bild, Ton, Video) gibt es hier mehr oder weniger robuste Ansätze, wobei die Weiterentwicklung von generativer KI eben, wie oben ausgeführt, sehr dynamisch ist und Verfahren, die sich heute als robust darstellen in Zukunft möglicherweise nicht mehr aussagekräftig sein können.

(1) Artefakte im Produkt

Beim Einsatz dieser Erkennungsmethoden ist sicherzustellen, dass sie zuverlässig funktionieren. Denn es wurde nachgewiesen, dass z. B. Texte, die von Personen in einer Fremdsprache verfasst wurden, die Treffgenauigkeit von KI-Detektoren beeinträchtigt, da sie häufiger fälschlicherweise als KI-generiert eingestuft werden (Liang et al. 2023). Die Objektivität der automatisierten Text-Detektoren kann hier also nicht gewährleistet werden, womit es zu unfairen Anschuldigungen wegen einer nicht offengelegten Nutzung kommen kann. Gerade in Bildungseinrichtungen mit einer hohen Anzahl von nicht-englischsprachigen Sprecher:innen sollte man sich deshalb nicht ausschließlich auf technische Detektoren verlassen, die die Echtheit von englischen Texten bewerten sollen. Ein Lösungsansatz hierfür könnten internationale Standards für Erkennungsmethoden sein. Diese könnten einerseits eine Evaluierung derartiger Verzerrungen beinhalten, sie würden aber auch zur Vereinfachung der Durchsetzung von KI-Detektion beitragen und praktische Leitlinien für Technologieentwickler:innen bieten Vasse'i/Udoh (2024).

(2) Fingerprints oder Wasserzeichen schon bei der Herstellung

Als zweiten Ansatz gibt es die Möglichkeit, bei Erstellung von Inhalten mittels sog. Fingerprints oder Wasserzeichen festzuhalten, dass diese entweder generative KI als Ursprung haben, oder aber z. B. mit einem Fotoapparat aufgenommen wurden. Hierzu bietet Abschnitt 2.3.1 die technischen Details. Auch hier gibt es gesellschaftspolitische Problemstellungen, die über die technische Ebene hinausgehen. So gibt es Bedenken gegenüber der Rückverfolgbarkeit von generierten Inhalten und der eindeutigen Identifizierung der Ersteller:innen durch die Einführung maschinenlesbarer Wasserzeichen. Dies könnte zur Verletzung von grundlegenden Menschenrechten, wie dem Datenschutz, führen. Beispielsweise könnten Dissident:innen oder Whistleblower identifiziert und verfolgt werden. Grundsätzlich sollte nicht nur die technische Machbarkeit im Vordergrund stehen, sondern auch die Frage, ob dadurch in bestimmten Fällen auch Menschenrechte verletzt werden können (van Huijstee et al. 2021). In einem Diskussionspapier wird deshalb von einer verpflichtenden Kennzeichnung durch kryptografische Wasserzeichen, die zur Identifizierung von Personen direkt oder indirekt über individuellen Eingaben geeignet sind, abgeraten (Björksten 2023, S. 10ff). Damit ist klar, dass auch bei diesen Ansätzen oft ein Abwägen zwischen sinnvoller Transparenz und ausgeweiteter Überwachung notwendig ist (Vasse'i/Udoh 2024).

Open-Source-Wasserzeichen

In engem Zusammenhang mit der Zugänglichkeit von Erkennungsmethoden wird deshalb empfohlen, die Durchführbarkeit von Open-Source-Wasserzeichen- und Erkennungsmethoden mit Nachdruck zu untersuchen. Dies könnte zu innovativeren und leichter zugänglichen Methoden führen, um die Authentizität von Inhalten zu gewährleisten (Vasse'i/Udoh 2024).

6.3.2 TECHNIKEN GEGEN FALSCHINFORMATION

Während Generative KI besonders als Werkzeug zur Diskursverzerrung mittels Falschinformation demokratiepolitisch problematisch sein kann, liefert Generative KI nicht jedes Mal Falschinformation, noch ist jede Falschinformation mit Generativer KI erstellt. Daher lohnt es sich, auch abseits der Frage, wie Falschinformationen erstellt wurden, technische Ansätze zu beleuchten, die dieser entgegengesetzt werden können. Dabei wollen wir drei Ansätze hervorheben: Erstens gibt es verschiedene Ansätze, um per Software Falschinformation zu erkennen bzw. Menschen die Erkennung zu erleichtern. Zweitens liegt ein Augenmerk oft auf den Verbreitungsmustern von Falschinformation. Und drittens gibt es verschiedene Möglichkeiten, um deren Verbreitung einzudämmen – hier kommt besonders Online-Plattformen eine zentrale Bedeutung zu.

Open-Source Intelligence

Bei der Überprüfung der Echtheit von Text-, Bild- und Videomaterial gewinnen die Stärkung von Basiswissen der Medienschaffenden und der Konsument:innen über Open-Source Intelligence (OSINT) und die Entwicklung KI-gestützter, automatisierter Fake-Erkennung an Bedeutung.[197] Unter OSINT wird hier das Sammeln und Analysieren von Informationen aus offenen, online verfügbaren Daten und Quellen verstanden, um daraus nützliche Erkenntnisse zu gewinnen.[198] Ein Einsatz dafür ist, Rückschlüsse auf die Echtheit von Informationen wie Bildern und Videos zu machen. Während der Begriff OSINT ursp. aus der Welt der Nachrichtendienste stammt, kann die Methode im hiesigen Kontext als eine Art Faktencheck zur Verifizierung von Inhalten (im politischen Diskurs) verstanden werden. Das wird von verschiedenen Institutionen bereits eingesetzt, vom Bereich der Nationalen Sicherheit und Strafverfolgung bis hin zu NGOs, am bekanntesten ist hier vielleicht Bellingcat[199] (Karaboga et al. 2024, S. 234).

Es wurden verschiedene Softwarelösungen entwickelt, um aus offenen Quellen Daten zu sammeln, zu analysieren und zu verwerten. Mit diesen Hilfsmitteln können Benutzer:innen effizient Informationen aus Datenbanken extrahieren und dadurch Hinweise auf die Vertrauenswürdigkeit erhalten. Beispiele sind: Maltego, Shodan, TheHarvester,[200] die Nutzung von Google Dorks,[201] oder Recon-NG.[202]

Diese Methoden und ähnliche technische Hilfsmittel für Endverbraucher:innen verlangen, im Gegensatz zur maschinenlesbaren Kennzeichnung, hohe Kompetenz seitens der Anwender:innen. Außerdem bergen sie das Risiko, dass die Verantwortung auf die Endnutzer:innen übertragen wird, die bereits mit zu vielen Informationen überfordert sein können (Vasse'i/Udoh 2024).

---

197 Z. B. *lawfaremedia.org/article/finding-language-models-in-influence-operations#:%7E:text=Approach%202%3A%20Automated%20Detection%20of%20AI%2DGenerated%20Text*; *cordis.europa.eu/project/id/101070093*.

198 *data.europa.eu/en/publications/datastories/open-source-intelligence*.

199 *bellingcat.com*.

200 *maltego.com*; *shodan.io*; *kali.org/tools/theharvester/*.

201 Google Dorks (engl. für "Deppen") sind fortgeschrittene Nutzungsmöglichkeiten der Google-Suche, siehe z. B. *recordedfuture.com/threat-intelligence-101/threat-analysis-techniques/google-dorks*.

202 Recon-ng ist eine auf Informationsbeschaffung zugeschnittene Umgebung, die das Finden, Sammeln und Organisieren von OSINT-Daten über verschiedene Schnittstellen unterstützt. Siehe z. B. *biteno.com/was-ist-recon-ng/*.

Browser-Erweiterungen zur Analyse und Kennzeichnung

Weiterhin werden Browser-Erweiterungen zur Analyse und Kennzeichnung von Nachrichten auf Webseiten[203] angeboten, denen journalistische Kriterien zugrunde liegen (z. B. NewsGuard[204]). Es werden aber auch Systeme angeboten, die Webseiten mit bekannten Seiten abgleichen, bei denen Desinformation bereits nachgewiesen wurden (Decodex[205] war z. B. so ein Projekt). Die Informationen beruhen oftmals auf manuellen Bewertungen durch Expert:innen und sind deshalb nur begrenzt einsetzbar. Die Identifizierung von Desinformation wird deshalb auch mit KI-Methoden unterstützt, die auf maschinellem Lernen oder auch auf sogar selbst auf Sprachmodellen beruhen (Barmann et al. 2024). Die Kombination von Identifizierung und Kennzeichnung von Desinformation als Browser-Plug-in aufgrund charakteristischer Stilmittel sind auch Gegenstand vieler Forschungsprojekte. Ein Beispiel stellt das Projekt „Desinformationskampagnen beheben durch Offenlegung der Faktoren und Stilmittel" (DeFaktS) dar.[206]

Analyse der Verbreitungsmuster

Oft werden zur Erkennung von Desinformation auch deren Verbreitungsmuster analysiert. So erforscht das Forschungsprojekt „DESINformations Früherkennung von gefährdenden online nAChrichten Trends" (Desinfact[207]) die Erkennung von Desinformationskampagnen sowohl durch Analyse der Inhalte als auch der Netzwerkstruktur und Kommunikationsmuster. In eine ähnliche Richtung ging das Projekt PHEME,[208] das sich besonders auf virale Nachrichten und deren Wahrheitsgehalt fokussierte.

Maßnahmen der Plattformen

Bei der (Weiter-)Verbreitung von Deepfakes und anderen Formen von Falschinformationen kommen Online-Plattformen besondere Bedeutung zu, besonders jene mit großer Nutzer:innenbasis. Ein wichtiger Baustein gegen Falschinformationen sind daher jedenfalls Maßnahmen dieser Plattformen, um die Verbreitung von Deepfakes zu erschweren. Als Beispiel in diese Richtung kann Google dienen. Der Suchmaschinenanbieter reiht Webseiten, die in der Vergangenheit Deepfakes veröffentlicht haben, besonders jene mit sexuell explizitem Inhalt, weiter hinten in den Suchergebnissen, sie werden also als wenig relevant herabgestuft – und somit weniger Nutzer:innen angezeigt.[209]

---

203 Siehe dazu die Verpflichtung für große Plattformen und Suchmaschinen in Art 35 (1) k DSA: „Sicherstellung, dass eine Einzelinformation, unabhängig davon, ob es sich um einen erzeugten oder manipulierten Bild-, Ton- oder Videoinhalt handelt, der bestehenden Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähnelt und einer Person fälschlicherweise als echt oder wahrheitsgemäß erscheint, durch eine auffällige Kennzeichnung erkennbar ist, wenn sie auf ihren Online-Schnittstellen angezeigt wird, und darüber hinaus *Bereitstellung einer benutzerfreundlichen Funktion, die es den Nutzern des Dienstes ermöglicht, solche Informationen anzuzeigen*." (unsere Herv.)

204 *newsguardtech.com*.

205 *lemonde.fr/les-decodeurs/article/2017/02/03/decodex-notre-kit-pour-verifier-l-information-a-destination-des-enseignants-et-des-autres_5074257_4355770.html*.

206 *neueshandeln.de/sprich/factchecking-und-ki-chancen*; *fzi.de/project/defakts/*.

207 *donau-uni.ac.at/de/forschung/projekt/U7_PROJEKT_4294971093*.

208 *pheme.eu*.

209 *blog.google/products/search/google-search-explicit-deep-fake-content-update/*.

6.3.3 NEUE ONLINE-PLATTFORMEN

Value sensitive design

Im Abschnitt 6.2.2 wurden bereits nicht-kommerzielle, demokratieverträgliche Soziale Medien genannt, die sich nicht an Profitmaximierung orientieren, sondern stattdessen einen demokratieverträglichen Diskursraum schaffen. Daran knüpfen Vorschläge für ein „value sensitive design" von Plattformen an, die mehr auf die Unterstützung und das Engagement ihrer Nutzer:innen setzen als auf das möglichst lange Verweilen auf ihren Plattformen (EGE 2023, S. 25).

Relevanz ohne Personalisierung

Darüber hinaus braucht es aber auch neue Ansätze, um epistemische Bubbles, Polarisierung und Misstrauen vorzubeugen. Es müssten Verfahren entwickelt werden, um Nutzer:innen relevante Informationen anzuzeigen, ohne auf Personalisierung zurückzugreifen – und damit Gefahr zu laufen, einen zunehmend fragmentierten öffentlichen Diskurs zu verstärken.

Positives vor- statt Negatives nachreihen

Eine weitere Möglichkeit wäre, Inhalte gezielt zu priorisieren, die z. B. besonders diskursförderlich, vertrauenswürdig, multiperspektivisch, oder kurzum: demokratieförderlich sind – quasi als positives Gegenstück zum Herabstufen von Deepfakes und Desinformation im vorigen Abschnitt. Über Inhalte hinausgehend gibt es noch viel Forschungsbedarf, welche technischen Mechanismen und Interaktionsmuster einen fruchtbaren, konstruktiven und für alle Involvierten bereichernden Online-Diskurs ermöglichen – einige Projekte rund um Bürger:innenbeteiligung versuchen, dies zu realisieren (siehe Abschnitt 3.3).

Forschungsbedarf

6.3.4 SLOW AI UND ZUVERLÄSSIGE KI

Statt Wettlauf der führenden KI-Unternehmen …

Viele der KI-Lösungen, die derzeit im Zentrum der Aufmerksamkeit stehen, fokussieren sich darauf, immer „besser" zu werden und breitere Schichten an Nutzer:innen anzusprechen. Es gibt eine gewisse Erwartungshaltung, KI müsse eingesetzt werden, um nicht „abgehängt" zu werden. Besonders seit der Veröffentlichung von ChatGPT Ende 2022 findet ein sehr intensiver Wettbewerb unter den führenden Unternehmen um Marktanteile statt. Wachstum, Marktbeherrschung, Akzeptanz und technische Überlegenheit scheinen hier derzeit die Entwicklung anzutreiben (siehe Abschnitt 4.4).

… Systeme testen, bevor sie auf den Markt kommen

Als Gegenbewegung dazu gibt es auch viele Stimmen, die eine andere Prioritätensetzung fordern. In der Mozilla-Studie plädiert man zum Beispiel für die Förderung und Bereitstellung von mehr F&E und Finanzmitteln für „langsame" KI-Lösungen (abgeleitet von „slow tech"), die die soziale Verantwortung der Unternehmen in die Technologie einbeziehen (Vasse'i/Udoh 2024). So könnten beispielsweise Kennzeichnungs- und Erkennungssysteme für generierte Inhalte auf ihre Wirksamkeit getestet werden, *bevor* ein KI-System eingeführt wird.

… und Ethics Guidelines for Trustworthy AI

In eine ähnliche Stoßrichtung gehen auch verschiedene Versuche zu definieren, was vertrauenswürdige KI auszeichnet und wie die Entwicklung dieser sichergestellt bzw. gefördert werden kann. Paradigmatisch dafür stehen die Ethics Guidelines for Trustworthy AI der High-Level Expert Group on AI (2019). Auch hier sind einige Grundsätze und Prinzipien festgehalten, die eine KI-Entwicklung anleiten möge, die zu robuster, nachhaltiger, demokratieverträglicher, fairer und den Menschen in den Mittelpunkt stellender KI führen soll – bei vielen Generative KI-Systemen könnte man argumentieren, dass sie einige dieser Grundsätze missachten.

7 SCHLUSSFOLGERUNGEN FÜR DAS ÖSTERREICHISCHE PARLAMENT

Vorgangsweise, um Schlussfolgerungen zu ziehen

In Kapitel 6 wurden die in der Rechtsetzung, in der Praxis und der wissenschaftlichen Literatur erörterten bzw. zur Anwendung gebrachten Handlungsansätze zusammengestellt und sortiert. Eine Auswahl jener Optionen, die bislang noch nicht (hinreichend) implementiert wurden und die aus Sicht des Projektteams geeignet scheinen, die aufgezeigten Herausforderungen für die Demokratie zu meistern, wurden in einem Workshop (siehe Anhang) mit Expert:innen und einigen politischen Vertreter:innen diskutiert und priorisiert; anschließend wurden die fünf am höchsten bewerteten Optionen einer SWOT-Analyse (siehe Anhang) unterzogen. Die Ergebnisse des Workshops wurden durch das Projektteam vertiefend analysiert sowie durch den interdisziplinären Projektbeirat und in einem Qualitätssicherungsseminar am ITA validiert. Das Analyseergebnis wurde in Kapitel 6 eingearbeitet. Die folgenden Schlussfolgerungen basieren auf diesem Verdichtungs- und Überarbeitungsprozess und stellen eine Auswahl der erfolgversprechendsten Optionen dar.[210] Sie richten sich primär an das österreichische Parlament als zentrale legislative Instanz, aber auch an die interessierte Öffentlichkeit im Sinne eines demokratischen Diskurses und als Grundlage der informierten Willensbildung (vgl. Vorwort).

Optionen-Mix notwendig, Regulierung alleine reicht nicht

Ein zentrales Ergebnis der Analyse der Handlungsoptionen ist, dass es keine einzelne Maßnahme zur Lösung einer, geschweige denn aller Herausforderungen zugleich gibt. Vielmehr ist immer eine *Kombination aus verschiedenen Ansätzen sowie eine systemische Betrachtung und Herangehensweise notwendig*. Insbesondere reicht trotz aller Notwendigkeit eines steuernden Rechtsrahmens Regulierung alleine kaum aus. Die Gründe dafür sind vor allem die Schwierigkeit bei der Rechtsdurchsetzung im internationalen Raum, nicht nur für den Nationalstaat, sondern sogar für die supranationale EU, sowie die Abwägung unterschiedlicher auch teils konfligierender Grundrechte wie Meinungs- und Kunstfreiheit gegenüber der staatlichen Pflicht zum Schutz von Persönlichkeitsrechten. Weiterhin haben wir es mit sehr unterschiedlichen Akteuren zu tun – von der Erstellung, der Verbreitung bis zur Rezeption der KI-generierten Inhalte. Erschwerend kommt hinzu, dass klassische Gatekeeper für vertrauenswürdige Informationen im Bereich der Medien eine immer schwächere Rolle spielen und damit die Bürger:innen wie auch Entscheidungsträger:innen mit einer Flut an Informationen unterschiedlicher Qualität konfrontiert sind. Dazu kommt die Langwierigkeit der Rechtswege, wenn Akteure, die sich nicht an die Regeln halten, geklagt werden (denn die Erfahrung zeigt, dass praktisch immer Einspruch erhoben wird). Die Rechtsetzung in Demokratien ist darüber hinaus (aus guten Gründen) relativ schwerfällig und kann oftmals nur zeitverzögert auf die äußerst dynamische Entwick-

[210] Wir verzichten aus Gründen der Lesbarkeit in diesem Kapitel auf Quellenverweise. Alle hier genannten Vorschläge wurden in der einen oder anderen Weise mit entsprechenden Nachweisen in Kapitel 6 vorgestellt.

lung der Technik und des Marktes reagieren. Daher ist auch das bisher Beschlossene (v. a. auf EU-Ebene) zwar ein wichtiger erster Schritt, aber allein kaum in der Lage, den Diskursraum für eine gedeihliche Demokratie sicherzustellen. Aus diesem Grund erscheint es essentiell, Initiativen proaktiver Art in verschiedenen Dimensionen zu setzen, also sowohl regulative Maßnahmen zu setzen als auch diese durch organisatorische und technische zu flankieren.

DEN DEMOKRATISCHEN DISKURS STÄRKEN

Enquetekommission „Demokratie und KI“

Als Basis-Maßnahme wird hier vorgeschlagen, dass sich die zentralen politischen Akteure in Österreich, insbesondere die Mandatar:innen auf Bundes- und Landesebene, weitere Vertreter:innen von Parteien, der Sozialpartner sowie Expert:innen aus Zivilgesellschaft und Wissenschaft *in Form einer parlamentarischen Enquetekommission oder einer sonstigen hochrangigen Veranstaltungsreihe über die Zukunft des politischen Diskurses* in Zeiten von Social Media und KI austauschen und verständigen. Neben einem Anstoß für eine reflektierte politische Praxis und für eine gesellschaftsweite Auseinandersetzung mit dem Thema selbst könnten dabei auch *Impulse für Selbstregulierung* gesetzt werden (siehe dazu spezifische Vorschläge in Kapitel 6). Zentral wäre die Erarbeitung von quer zu allen politischen Richtungen konsensfähigen Mechanismen, um zwischen dem Grundrecht auf freie Meinungsäußerung inklusive legitimer politischer Kampagnisierung einerseits und dem nicht minder wichtigen Erhalt eines faktenbasierten demokratischen Diskurses auszugleichen.

Verhaltenskodex KI in der Politik

Im Zeichen verantwortungsvoller Politik könnte ein möglicher Konsens in diesen Bereichen letztlich zu einem *Verhaltenskodex für Politiker:innen* mit Regeln über den Gebrauch von Generativer KI im politischen Geschäft inklusive Wahlkämpfen verdichtet werden. So ein freiwilliger Verhaltenskodex könnte dadurch verbindlicher gestaltet werden, dass eine Art *Digitaler Ordnungsruf* eingeführt wird, um bei Missachtung der vereinbarten Regeln für den Umgang mit Sozialen Medien ähnlich den Ordnungsrufen im Parlamentsplenum durch den/die Parlamentspräsident:in zu sanktionieren. So ein Ordnungsruf könnte in der Geschäftsordnung des Parlaments verankert werden oder als Aufgabe einer sonstigen unabhängigen Stelle eingerichtet werden, ähnlich dem Presserat, also einer Art *Ethikrat für politische Werbung, KI und Public Relations in Sozialen Medien*.

Bürger:innen-Foren zu Grundsatzfragen der Demokratie

Es braucht aber nicht nur innerhalb der Politik, sondern auch in der Öffentlichkeit eine umfassende Debatte darüber, was Demokratie bzw. demokratischer Diskurs bedeuten soll und wie sich unsere Gesellschaft die Zukunft diesbezüglich vorstellt. Die Autor:innen dieser Studie sind der Überzeugung, dass dazu grundsätzliche und breit geführte Debatten – über das Parlament hinaus – essentiell sind. Es wird zunächst einen Grundkonsens über die Art von Demokratie, die Österreich ausmachen soll, benötigt. Eine Option wäre eine *Serie von Bürger:innen-Foren zur Konsensfindung in Grundsatzfragen der demokratischen Debattenkultur*, beginnend im Parlament (siehe oben: Vorschlag Enquetekommission), in weiterer Folge mit vielen Veranstaltungen quer durch Österreich, auch mit medialer Begleitung. Die Foren müssten so konzipiert werden, dass ein wertschätzender und produktiver Dialog möglich ist. In diesem Rahmen zu erörternde und auf Basis eines entsprechenden Moderationskonzepts auf die Alltagswahrnehmungen der Bürger:innen anzupassende Fragen könnten unter anderem folgende sein: Welche Debattenkultur braucht es für eine funktionierende Demokratie? Was ist Desinformation und wie und auf welcher Grundlage soll und kann ent-

schieden werden, was im demokratischen Diskurs außer Streit gestellt werden soll? Wie kann behördliche von parteipolitischer Kommunikation getrennt werden? Wie kann die Moderation von Inhalten im Zeitalter sozialer Medien gestaltet werden, um sowohl Freiheits- wie Persönlichkeitsrechte zu wahren? Welche Rolle können neuartige Technologien wie Generative KI in Deliberationsprozessen sowie zur Erarbeitung von innovativen Lösungen komplexer Problemstellungen spielen?

Förderung von Medien- und KI-Literacy

Ein zentrales Ergebnis dieser Studie ist, dass unreflektierte aktive Nutzung der neuen Tools, aber auch kritikloses passives Konsumieren bei entsprechend weiter Verbreitung eine Gefahr für den informierten, freien Willensbildungsprozess als zentralen Grundpfeiler der Demokratie darstellt. Daher wäre die nach Auffassung der meisten Expert:innen wichtigste Maßnahme eine *gezielte Förderung des Bewusstseins in der allgemeinen Bevölkerung* bezüglich Generativer KI. Das schließt auch den Wissenserwerb zu KI und Generativer KI sowie zu traditionellen und sozialen Medien in der Bevölkerung ein. Da solche Maßnahmen jedoch erst mittel- bis langfristig wirken, sind sie *besonders dringlich, um rechtzeitig wirksam werden zu können*, müssen aber gleichzeitig kontinuierlich und zielgruppenspezifisch umgesetzt werden. Diese Maßnahmen sollten nicht zu eng auf den Missbrauch der Technologie, sondern insbesondere auf Stärkung der unabhängigen Institutionen der Demokratie und die Resilienz der Bürger:innen ausgerichtet werden (also auf staatliche und individuelle digitale Souveränität). Dazu müssten in Schule und Erwachsenenbildung entsprechende Module erarbeitet werden, die im besten Fall interaktives Lernen ermöglichen (etwa zu den Themen: Wie funktioniert Generative KI? Welche Kennzeichnung ist hilfreich? Welche Risiken und Gefahren sind mit den durch Generative KI produzierten Inhalten verbunden? Wie funktioniert Fact-Checking und wo sind vertrauenswürdige Faktenchecks zu finden?). In der Schule könnte der Lehrplan entsprechend überarbeitet werden und diese Thematik als Teil einer intensivierten politischen Bildung (vor Erreichen des gesetzlichen Wahlalters) konzipiert werden. Für Erwachsene müssten weitere Kanäle der Vermittlung dieses Wissens genutzt werden (Plakate, Aufrufe zur Teilnahme an Schulungen, online oder in den Volkshochschulen, Belangsendungen, Anzeigenschaltung im öffentlichen Interesse in Medien aller Art usw.). Angesichts der Vielsprachigkeit der in Österreich lebenden Bevölkerung wären diese Angebote am besten nicht nur auf Deutsch zu machen und müssten auch bildungsferne Personen ansprechen können. Da Desinformation und Manipulation insbesondere in Gesellschaften mit geringem Institutionenvertrauen wirken, wird weiters vorgeschlagen, besonders die Rolle von zivilgesellschaftlichen Organisationen zu stärken, die einerseits vielfach bereits umfangreiches relevantes Wissen aufgebaut haben, andererseits auch staatliche Stellen als Watchdog-Organisationen zu stärken. So kann der öffentliche Diskurs mit unterschiedlichen Wissensbeständen und Perspektiven angereichert werden, die auch eine für Demokratie essentielle Objektivierungsfunktion erfüllen.

Transparenzerhöhung: Kennzeichnung und Ansätze zu Erklärbarkeit

Um das Potenzial von KI auszuschöpfen und eine verantwortungsvolle und sozialverträgliche Gestaltung voranzubringen, erscheint es insbesondere im Zusammenhang mit einer Bildungsoffensive wichtig, die Transparenz im Zusammenhang mit KI generell und Generativer KI im Speziellen zu erhöhen. Dazu sollten verschiedene Maßnahmen zur *umfassenden Kenntlichmachung* von KI-generierten Inhalten ebenso ergriffen werden wie auch Ansätze zu sog. Erklärbaren KI (explainable artificial intelligence, XAI) zu fördern.

Spezifische Regulierungsoptionen

Auch wenn, wie oben argumentiert, Regulierung alleine nicht in der Lage ist, die Problematik der Diskursverzerrungen durch Generative KI im politischen Feld zu entschärfen, so gibt es doch eine Reihe von Vorschlägen, die Teil der Lösung sein könnten und daher dem Gesetzgeber nahegelegt werden. Dabei ist zu beachten, dass diese rechtssetzenden Aktivitäten in der Regel auf EU-Ebene stattfinden müssten, der Beitrag Österreichs also im Wesentlichen darin bestehen würde, konstruktiver und engagierter Akteur im Mehrebenensystem zu sein. Beispiele für potenzielle regulative Maßnahmen, die über die bestehende Rechtslage hinausgehen, sind:

- Online-Plattformen als (teil-)verantwortlich für die auf ihnen und durch sie (algorithmisch gesteuert) verbreiteten Inhalte einstufen, ähnlich zu Medien. Damit einher geht die Verpflichtung, wirksame Mechanismen zu entwickeln, um rechtswidrige Inhalte und/oder Desinformation unter Wahrung der Grundrechte zu löschen.
- Klare rechtliche Grenzen für Microtargeting in der Politik und Achtung ethischer Grundsätze;
- Verbot von Deepfakes politischer Kandidat:innen, welche diesen in der politischen Auseinandersetzung schaden sollen;
- Generelle Vorschrift, dass Suchmaschinen Ergebnisse von jenen Quellen, die nachweislich vielfach Desinformation oder generierte Inhalte verbreiten, in den Suchergebnissen nachgereiht werden sollen;
- Vorschrift, dass durch Generative KI gesteuerte Bots ihre Kommunikationspartner:innen nicht im Unklaren über ihre Künstlichkeit lassen dürfen, also explizit als Bots gekennzeichnet sind, weil sonst die Authentizität und Integrität von Kommunikation untergraben würde.

DIGITALE SOUVERÄNITÄT AUF DEMOKRATISCHEN PRINZIPIEN AUFBAUEN

Österreich als Vorreiter einer proaktiven Vorgangsweise zur Resilienzerhöhung der Demokratie

Es ist offensichtlich, dass es angesichts der internationalen Vernetzung nicht effizient sein kann, wenn ein Nationalstaat alleine aktiv wird. Die supranationale Ebene kann aber nur so stark sein, wie ihre Mitglieder es wollen. Daher braucht es mutige Vorschläge seitens der Nationalstaaten. Österreich könnte vom passiv nachvollziehenden Politiknehmer auf EU-Ebene zu einem starken Befürworter und aktiven Impulsgeber eines geregelten und fairen Umgangs im demokratischen Diskursraum wandeln. Österreich könnte sich auf Basis entsprechender innerstaatlicher Vorarbeiten (siehe die Gesamtheit der vorgestellten Optionen in diesem Kapitel) konsequent mit Initiativen im Rechtssetzungs- und Politikformulierungsprozess innerhalb der EU einbringen und das wichtige Thema „Erhalt der Demokratie europäischer Prägung“ voranbringen. Dazu braucht es ein starkes Commitment, das von vielen getragen wird, und einen langen Atem. Dabei geht es nicht nur um den Erhalt im Inneren, sondern auch um die Abwehr von Bedrohungen von außen, befinden sich die Demokratien doch in einem sich verschärfenden Systemwettbewerb mit Autokratien.

Initiativen in Richtung staatliche digitale Souveränität

Vor dem Hintergrund, dass Technologieentwickler wie Plattformunternehmen nicht demokratisch legitimiert sind, sollten Österreich und die EU ihre Bemühungen verstärken, digital souverän zu werden, um sich aus der aktuell massiven Abhängigkeit von nicht-europäischen Technologieanbietern zu lösen und damit mehr Handlungsspielraum zu gewinnen. Dies inkludiert eine Wirtschafts-, Standort-, Förderungs- und Innovationspolitik, die langfristig die entsprechenden *digitalen Infrastrukturen* auch in Österreich und Europa aufbaut, inklusive jener, die politiknahe und somit Teil der „demokratischen Daseinsvorsorge“ sind (res-

ponsive Verwaltung, unbeeinflussbares Wahlsystem, qualitätsvolle Diskursräume in Medien und auf Plattformen etc.). Dazu braucht es ein starkes Europa, um den weltweiten Entwicklungen standhalten bzw. etwas entgegensetzen zu können. Zentral könnte die staatliche/EU-weite Schaffung von Rahmenbedingungen sein, die für Start-ups und bereits am Markt befindlichen Firmen förderlich sind. So müssten etwa das *Kartellrecht und die Regeln gegen unlauteren Wettbewerb* streng zur Anwendung gebracht werden, um das Aufkommen dieses Unternehmenszweiges in Europa angesichts der zum Teil bereits marktbeherrschenden Stellung nicht-europäischer Anbieter auszugleichen. Weiters können F&E-Initiativen KI-Anwendungen, die in Österreich und in Europa entwickelt werden, verstärkt werden. Dabei sollten auch Anreize für eine Fokussierung auf *demokratieverträgliche KI* gesetzt werden – Alternativen zu US- und China-dominierter KI könnten hier auch neue Impulse und Möglichkeiten für europäische Unternehmen bieten. Aus demokratiepolitischer Sicht wären gerade auch Incentives für die Entwicklung von *Debunking-Software* zur Unterstützung von Fact-Checking bei der Detektion von Inhalten, die durch Generative KI erzeugt wurden, besonders wichtig. Auch die Förderung von *Open Public Data* in Hinblick auf die Verfügbarkeit von hochwertigen, *nicht-englischsprachigen* und kulturell vielfältigen Trainingsdaten in Österreich/Europa wäre ein wichtiger Beitrag in Richtung digitaler Souveränität.

Demokratieverträgliche europäische Diskursplattform

Da Generative KI zunehmend auch in den sozialen Medien zum Einsatz kommt, könnte es ein spezifisches Ziel in diesem Zusammenhang sein, Europa eine oder mehrere von rein kommerziellen und nicht-europäischen Interessen unabhängige, qualitätsgesicherte, datenschutzkonforme, rechtskonform agierende und vor allem proaktiv Hassrede und Desinformation hintanhaltende und dem Gemeinwohl verpflichtete, sprich: demokratieverträgliche Diskursplattformen aufzubauen – sozusagen ein echtes *soziales* Medium, um einen breiten Diskursraum zu schaffen. Das könnte dazu führen, dass Europa – und damit auch Österreich – nicht mehr von den Strategien nicht-europäischer Unternehmen im Social-Media-Bereich abhängig ist, wenn es um die Bereitstellung einer zentralen Infrastruktur für die Demokratie geht. Während One-to-many-Kommunikation (etwa über die klassischen Medien, Plakate etc.), aber auch Few-to-few-Kommunikation (innerhalb von Parteien, am klassischen Stammtisch, in der Familie) zwar erhalten bleiben werden, kann davon ausgegangen werden, dass elektronische Many-to-many-Kommunikation eine immer wichtigere Funktion im politischen System spielen wird. Während es für die hergebrachten Kommunikationsformen etablierte Spielregeln gibt, scheint es essentiell, auch für die neuen Many-to-many-Formen demokratiefördernde Infrastrukturen zu schaffen. Wie dies am besten zu realisieren ist, muss an dieser Stelle offenbleiben. Es gibt verschiedene Optionen, von der massiven Förderung privater Initiativen über Public-Private-Partnerships bis zu öffentlich-rechtlichen Modellen. Die große Herausforderung wird es sein, diese neue(n) Plattform(en) so attraktiv zu gestalten, dass sie von den Europäer:innen tatsächlich angenommen werden. Die Frage, wie solche Plattformen gestaltet werden können, kann auch als Chance genutzt werden, um eine breite Diskussion über demokratische Debattenkulturen (siehe oben) zu führen.

Systematische und konsequente Abwehr von Manipulations-versuchen:

Zur systematischen und konsequenten Abwehr von Manipulationsversuchen werden die folgenden Maßnahmen vorgeschlagen:

- Öffentliche (finanzielle) Förderung für unabhängiges Fact-Checking, um es aus dem idealistischen, unterfinanzierten Bereich herauszuholen und als staatlich anerkannte, demokratieförderliche Maßnahme zu positionieren und zu skalieren, sei es über die generelle Medienförderung z. B. als von allen Medien getragene Organisation, sei es auf Basis von Anträgen von Fact-Checking-Akteuren; *Fact-Checking fördern*
- Vertiefung existierender Koordinationseinrichtungen zur Abwehr externer Einmischung in den innerstaatlichen Diskurs von außen bzw. Intensivierung entsprechender Bemühungen im Rahmen der nationalen Cyberstrategie im Austausch mit entsprechenden EU-Initiativen; *Koordination gegen externe Einmischung intensivieren*
- Förderung der Verwendung von Wasserzeichen und Fingerprints in Österreich/Europa, inkl. Browser bzw. Browserplugins zur Sichtbarmachtung der solcherart gekennzeichneten Inhalte; *Transparenz erhöhen*
- Durchsetzung einer umfassende Kennzeichnung KI-generierter Inhalte, *unabhängig* von der Größe der Plattform. *Kennzeichnung & Durchsetzung*

CHANCEN AUSLOTEN, BEGLEITFORSCHUNG FÖRDERN

Förderung chancenreicher KI-Anwendungen im politischen Kontext

Nicht aus den Augen verloren werden sollte, dass Generative KI nicht ausschließlich eine Bedrohung für die Demokratie sein muss, sondern auch das Potenzial hat, demokratieförderlich eingesetzt zu werden. Daher sollte die gezielte Förderung der (Weiter-)Entwicklung der vielversprechenden Anwendungen Generativer KI im politischen Kontext (siehe Kapitel 3) Teil des Maßnahmenpakets sein. Hier besteht nach Einschätzung der Studienautor:innen Potenzial, den politischen Diskurs zu verbessern, etwa dadurch, dass Fachdokumente für Bürger:innen und Politiker:innen besser zugänglich gemacht, zusammengefasst oder übersetzt werden, schriftliche Debatten „intelligent" strukturiert werden u. v. m. Dabei ist aber wichtig, die der Generativen KI inhärenten Nachteile wie fehlerhafte oder irreführende Inhalte mit effektiven Vorkehrungen Einhalt zu gebieten. Diese Förderung könnte einerseits über den o. g. Weg der F&E-Politik realisiert, andererseits als Aufgabe der staatlichen Rechenzentren (wie dem BRZ) definiert werden.

KI-Begleitforschung

Schließlich scheint es neben der bereits oben genannten F&E zu europäischen KI-Anwendungen angesichts der rasant zunehmenden Bedeutung von (Generativer) KI in unserer Gesellschaft und der hohen Dynamik des Sektors angezeigt, Forschung gerade auch zu gesellschaftlichen, kommunikations- und politikwissenschaftlichen, psychologischen und ethischen Fragestellungen im großen Stil zu fördern. Auch kontinuierliche Foresights und Technikfolgenabschätzungen zu spezifischen Anwendungsszenarien und zur Analyse von Risikopotenzialen wären dazu im europäischen Verbund zu priorisieren. Ein Element dieser Forschung wäre etwa ein regelmäßiger (jährlicher) Monitoringbericht zur digitalen politischen Kommunikation in Österreich. Die Studienautor:innen sind zur Überzeugung gelangt, dass die europäische Gesellschaft nur dann, wenn es gerade vor dem europäischen kulturellen Hintergrund umfassendes und ständig aktualisiertes Wissen in diesem Bereich gibt, den massiven Herausforderungen gewachsen sein wird.

Jährlicher Monitoringbericht zur digitalen politischen Kommunikation in Österreich

ANHANG: WORKSHOP

Am 25.11.2024 fand in den Räumlichkeiten des Parlaments (Lokal 3 – Hans Kelsen) ein mehrstündiger Workshop mit Expert:innen, Vertreter:innen politischer Fraktionen sowie Beobachter:innen aus der Parlamentsdirektion statt. Das Programm war wie folgt (siehe Box 5).

Session 1: Status Quo
10:30-11:00 Präsentation der zentralen Studienergebnisse
11:00-12:00 Feedbackdiskussionen an 4 thematischen Tischen in 3 Runden mit Gruppenwechsel

Session 2: Optionen
13:00-13:20 Einführung und Überblick
13:20-14:00 Plenardiskussion
14:00-14:30 Moderierte SWOT-Analyse[211] auf Plakaten durch die Teilnehmer:innen
14:30-15:00 Zusammenfassung und Ausblick

Box 5: Workshop-Programm

Abgesehen vom fünfköpfigen Moderationsteam des ITA und drei Beobachter:innen aus der Parlamentsdirektion nahmen am Workshop die folgenden elf Personen teil (siehe Tabelle 8):

Tabelle 8: Workshop-Teilnehmer:innen

Name	Organisation
Beaufort, Maren, Dr.in	Österreichische Akademie der Wissenschaften (ÖAW), Institut für vergleichende Medien- und Kommunikationsforschung (CMC)
Boyer, Martin, Dipl.-Ing.	Austrian Institute of Technology (AIT), Center for Digital Safety & Security
Deimek, Gerhard, Dipl.-Ing.	Abgeordneter zum NR (FPÖ)
Ebenberger, Stefan, Mag.	Internet Service Provider Austria (ISPA)
Kadic, Alma	Verein Mimikama zur Aufklärung über Internetmissbrauch
Krenn, Brigitte, Dr.in	Austrian Research Institute for Artificial Intelligence (OFAI)
Krickl, Julia, MA	Österreichisches Institut für angewandte Telekommunikation (ÖIAT) & Watchlist Internet
Prem, Erich, Dr.	Universität Wien, Philosophy of Media and Technology & Eutema Research Services
Reichstädter, Peter, Dipl.-Ing.	Parlamentsdirektion: Abteilung IT-Strategie
Schuh, Harald, Mag.	Abgeordneter zum NR (FPÖ)
Stomper-Rosam, Bettina, Dr.in	Grüner Klub

[211] SWOT steht für Strenghts, Weaknesses, Opportunities, and Threads, also für Stärken, Schwächen, Chancen und Risiken. In einem interaktiven-diskursiven Prozess werden Aussagen zum jeweiligen Thema (in diesem Fall: einer Handlungsoption) gesammelt und in einer Vierfelder-Matrix protokolliert. Das Ziel ist es, durch Inputs unterschiedlicher Expert:innen ein möglichst umfassendes Bild zu gewinnen.

ABKÜRZUNGSVERZEICHNIS

Abs. .....Absatz
AFP .....Agence France Presse
A.I.D.....Artificial Intelligence and the Shaping of Democracy (Projekt)
AIT .....Austrian Institute of Technology
APA .....Austrian Press Agency
Art. .....Artikel (in einem Rechtsakt)
AWS.....Amazon Web Services
BK.....Bundeskanzler:in
BM.....Bundesminister:in
BRZ .....Bundesrechenzentrum
B-VG .....(österreichisches) Bundes-Verfassungsgesetz
C2PA.....Coalition for Content Provenance and Authenticity
CAI.....Content Authenticity Initiative
CAIL .....Criticial Artificial Intelligence Literacy (Projekt)
CMC.....Institut für vergleichende Medien- und Kommunikationsforschung
DAN.....do anything now
DoS.....denial of service
DPA .....Deutsche Presseagentur
DSA.....Digital Services Act (der EU)
DSGVO.....Datenschutz-Grundverordnung (der EU)
EFCSN .....European Fact-Checking Standards Network
EPTA.....European Parliamentary Technology Assessment (network)
Erw. .....Erwägungsgrund (in EU-Rechtsakten)
et al. .....und andere (Autor:innen)
EPRS .....European Parliament Research Service
EU .....Europäische Union
EU-V .....Verträge zur Gründung der Europäischen Union
F&E .....Forschung und Entwicklung
ff .....fortfolgende (Seiten)
FAIR-AI.....Fostering Austria's Innovative Strength and Research Excellence in Artificial Intelligence (Projekt)
FORWIT .....Rat für Forschung, Wissenschaft, Innovation und Technologieentwicklung
FTA .....Foresight und Technikfolgenabschätzung
GADMO.....German-Austrian Digital Media Observatory
GPT .....Generative Pre-Trained Transformer
GPU .....graphical processing unit
Herv. .....Hervorhebung
i.d.R. .....in der Regel
IKT .....Informations- und Kommunikationstechnik
insb. .....insbesondere
ISPA .....Internet Service Provider Austria
ITA .....Institut für Technikfolgen-Abschätzung (Wien)
ITAS .....Institut für Technikfolgenabschätzung und Systemanalyse (Karlsruhe)
Kap. .....Kapitel
KDD .....Koordinator für Digitale Dienste (gemäß DSA)
KI.....Künstliche Intelligenz

KIT .....Karlsruhe Institut für Technologie
KI-VO .....Verordnung (EU) über Künstliche Intelligenz
KommAustria.....Kommunikationsbehörde Österreich (bei der RTR)
LLM .....large language model
LMU.....Ludwig-Maximilians-Universität München
NATO .....North Atlantic Treaty Organization
NGO.....Non-Governmental Organization
NTA .....Netzwerk Technikfolgenabschätzung
o.ä. .....oder ähnlich(es)
ÖAW.....Österreichische Akademie der Wissenschaften
OECD.....Organization of Economic Cooperation and Development
ÖIAT .....Österreichisches Institut für angewandte Telekommunikation
OFAI .....Austrian Research Institute for Artificial Intelligence
ORF .....Österreichischer Rundfunk und Fernsehen
RLHF .....Reinforcement Learning from Human Feedback
RTR .....Rundfunk- und Telekom-Regulierungs-GmbH
S. .....Seite
sog. .....so genannte
STAA .....Science, Technology Assessment, and Analytics
STOA .....Science and Technology Options Assessment
SWOT .....Strenghts, Weaknesses, Opportunities, Threads
TA .....Technikfolgenabschätzung
TAB .....Büro für Technikfolgenabschätzung beim Deutschen Bundestag (Berlin)
TA-Swiss .....Technologiefolgenabschätzung Schweiz (Bern)
TKG.....Telekommunikationsgesetz (Österreich)
u.a. .....unter anderem
u.ä.m. .....und ähnliches mehr
Übers. .....Übersetzung
USA.....United States of America
vgl. .....vergleiche
VWA .....Vorwissenschaftliche Arbeit
XAI.....explainable artificial intelligence
z. B. .....zum Beispiel

LITERATUR

Abadi, M., Agarwal, A. und Barham, P., 2015, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Preominary Whitepaper*, ***https://www.tensorflow.org/static/extras/tensorflow-whitepaper2015.pdf***.

Aïmeur, E., Amri, S. und Brassard, G., 2023, Fake news, disinformation and misinformation in social media: a review, *Social Network Analysis and Mining 13(1)*, 30. ***https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9910783/***.

Albrecht, S., 2024, ChatGPT als doppelte Herausforderung für die Wissenschaft: Eine Reflexion aus der Perspektive der Technikfolgenabschätzung: *KI:Text*: De Gruyter, 13-28, ***https://www.degruyter.com/document/doi/10.1515/9783111351490-003/html***.

Allea, 2021, *Fact or Fake? Tackling Science Disinformation. Discussion Paper, 5*, Berlin: ALLEA, ***https://doi.org/10.26356/fact-or-fake***.

Allhutter, D., 2019, Of 'Working Ontologists' and 'High-quality Human Components'. The Politics of Semantic Infrastructures, in: Vertesi, J. und Ribes, D. (Hg.): *DigitalSTS: A Field Guide for Science & Technology Studies*, Princeton and Oxford: Princeton University Press, 326-348.

Ananthaswamy, A., 2023, In AI, is bigger always better?, *Nature 615(7951)*, 202-205. ***http://www.ncbi.nlm.nih.gov/pubmed/36890378***.

Ananya, 2024, AI image generators often give racist and sexist results: can they be fixed?, *Nature 627(8005)*, 722-725. ***https://www.nature.com/articles/d41586-024-00674-9***.

APA (Austria Presse Agentur), 2023, *APA presents new AI strategy "APA Trusted AI"*; Letzte Aktualisierung: 09.03.2023, ***https://value-news.apa.at/index.html%3Fp=5948.html***.

Arguedas, A. R. und Simon, F. M., 2023, Automating democracy: Generative AI, journalism, and the future of democracy. ***https://ora.ox.ac.uk/objects/uuid:0965ad50-b55b-4591-8c3b-7be0c587d5e7/***.

Atleson, M., 2023, Chatbots, deepfakes, and voice clones: AI deception for sale, *Federal Trade Commission*. ***https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale***.

Baldassarre, M. T., Caivano, D., Fernandez Nieto, B., Gigante, D. und Ragone, A., 2023, The social impact of generative ai: An analysis on chatgpt, *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*.

Baldé, C. P., Ruediger Kuehr, Tales Yamamoto, Rosie McDonald, Elena D'Angelo, Shahana Althaf, Garam Bel, Otmar Deubzer, Elena Fernandez-Cubillo, Vanessa Forti, Vanessa Gray, Sunil Herat, Shunichi Honda, Giulia Iattoni, S., D., Khetriwal, Vittoria Luda di Cortemiglia, Yuliya Lobuntsova, Innocent Nnorom, Noémie Pralat und Wagner, M., 2024, *The Global E-Waste Monitor 2024*, Geneva/Bonn: United Nations Institute for Training and Research (UNITAR), International Telecommunication Union (ITU).

Bender, E. M. und Koller, A., 2020, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data, *ACL 2020*, 2020/07//, ***https://aclanthology.org/2020.acl-main.463***.

Bertelsmann Stiftung, 2023, *Glossar: Methoden zum Umgang mit digitaler Desinformation*, ***https://www.bertelsmann-stiftung.de/de/unsere-projekte/upgrade-democracy/projektnachrichten/glossar-methoden-zum-umgang-mit-digitaler-desinformation***.

Beuth, P., Buschek, C., Heber, M., Rosenbach, M. und Tanriverdi, H., 2024, Wie das Weltbild einer künstlichen Intelligenz entsteht, *SPIEGEL Plus*. ***https://www.wiso-net.de/document/SPPL__4fc4bb687d00b2cab24af42556d601530a1b4a40***.

Bieber, C., Heesen, J., Grunwald, A. und Rostalski, F., 2024, *KI im Superwahljahr – Generative KI im Umfeld demokratischer Prozesse*; Whitepaper aus der Plattform Lernende Systeme, 2024-07, München: Lernende Systeme – Die Plattform für Künstliche Intelligenz/acatech.

Björksten, G., 2023, *Identifying Generative AI Content: When and How Watermarking Can Help Uphold Human Rights. A Discussion Paper*, September 2023: Access Now.

Blick, 2023, FDP macht Wahlkampf mit den Klima-Klebern, *Blick*. ***https://www.blick.ch/politik/per-ki-generiertes-sujet-fdp-macht-wahlkampf-mit-den-klima-klebern-id18717769.html***.

BMI/BKA/BMEIA/BMJ/BMLV (Bundesministerium für Inneres – Bundeskanzleramt – Bundesministerium für europäische und internationale Angelegenheiten – Bundesministerium für Justiz – Bundesministerium für Landesverteidigung), 2022, *Aktionsplan Deepfake*, im Auftrag von: Österreichisches Parlament, Nr. III-740 der Beilagen XXVII. GP – Bericht – 02 Hauptdokument, Wien: BMI, ***https://www.parlament.gv.at/dokument/XXVII/III/740/imfname_1466378.pdf***.

Bogner, A., Decker, M., Nentwich, M. und Scherz, C. (Hg.), 2022, *Digitalisierung und die Zukunft der Demokratie. Beiträge aus der Technikfolgenabschätzung*; in Reihe: Gesellschaft-Technik-Umwelt. Neue Folge, Bd. 23, hg. v. ITAS, Berlin: Nomos, ***https://www.nomos-elibrary.de/10.5771/9783748928928.pdf?download_full_pdf=1***.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E. und Fowler, J. H., 2012, A 61-million-person experiment in social influence and political mobilization, *Nature (489)*, 295-298. ***https://www.nature.com/articles/nature11421***.

Breithut, J., 2023, (S+) Deepfakes: Wie TV-Moderatoren zu Werbemarionetten von Betrügern werden, *Der Spiegel*, 2023/09/24/, ***https://www.spiegel.de/netzwelt/deepfakes-wie-tv-moderatoren-zu-werbemarionetten-von-betruegern-werden-a-2232ebc0-1995-4470-9ecd-48d56503336d***.

Bresnihan, P. und Brodie, P., 2021, New extractive frontiers in Ireland and the moebius strip of wind/data, *Environment and planning e: Nature and space 4(4)*, 1645-1664.

Bresnihan, P. und Brodie, P., 2023, Data sinks, carbon services: Waste, storage and energy cultures on Ireland's peat bogs, *new media & society 25(2)*, 361-383.

Brodie, P., 2020, Climate extraction and supply chains of data, *Media, Culture & Society 42(7-8)*, 1095-1114. ***https://journals.sagepub.com/doi/abs/10.1177/0163443720904601***.

Brynjolfsson, E., Li, D. und Raymond, L. R., 2023, *Generative AI at work*, Working paper Nr. 31161: National Bureau of Economic Research, ***https://www.nber.org/papers/w31161***.

BSI (Bundesamt für Sicherheit in der Informationstechnik), 2024, Deepfakes – Gefahren und Gegenmaßnahmen, *Bundesamt für Sicherheit in der Informationstechnik*. ***https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes.html?nn=1009560***.

Burger, R. und Jaeger, M., 2017, Der aufgemotzte Tür-zu-Tür-Wahlkampf der CDU, *Frankfurter Allgemeine*, ***https://www.faz.net/aktuell/politik/wahl-in-nrw/cdu-macht-wahlkampf-mit-einer-app-15009334.html***.

Burghardt, K., 2024, KI-Textgeneratoren und der Anspruch auf Wahrheit, in: Schreiber, G. und Ohly, L. (Hg.): *KI:Text. Diskurse über KI-Textgeneratoren*, Berlin/Boston: De Gruyter, 419-434, ***https://doi.org/10.1515/9783111351490***.

Burkart, R., 2021, *Kommunikationswissenschaft. Grundlagen und Problemfelder einer interdisziplinären Sozialwissenschaft*, 6. Aufl., Wien: UTB; Böhlau Verlag.

Byman, D. L., Gao, C., Meserole, C. und Subrahmanian, V. S., 2023, *Deepfakes and International Conflict, Brookings Institution*, ***https://www.getabstract.com/en/summary/deepfakes-and-international-conflict/46825***.

Cabinet Office (National Intelligence and Security Committee security and intelligence of U.K. Parliament), 2018, *Intelligence and Security Committee: Annual report 2016-2017*; Corporate report, 23 July, London, ***https://www.gov.uk/government/publications/intelligence-and-security-committee-annual-report-2016-2017***.

Calma, J., 2024, Microsoft's AI obsession is jeopardizing its climate ambitions, *The Verge*, 15.05., ***https://www.theverge.com/2024/5/15/24157496/microsoft-ai-carbon-footprint-greenhouse-gas-emissions-grow-climate-pledge***.

Casero-Ripollés, A., Tuñón, J. und Bouza-García, L., 2023, The European approach to online disinformation: geopolitical and regulatory dissonance, *Humanities and Social Sciences Communications 10(1)*, 1-10. ***https://www.nature.com/articles/s41599-023-02179-8***.

Caunes, K. (Hg.), 2023, *Artifical Intelligence and Democratic Values 2022*, Washington DC: Center for AI and Digital Policy, ***https://www.caidp.org/reports/aidv-2022/***.

Chan, C., Ginosar, S., Zhou, T. und Efros, A. A., 2019, Everybody Dance Now: arXiv, ***https://arxiv.org/abs/1808.07371***.

Chesney, R. und Citron, D. K., 2018, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*; SSRN Scholarly Paper, 2018/07/14/, Rochester, NY: Social Science Research Network, ***https://papers.ssrn.com/abstract=3213954***.

Chowdhury, R., 2024, AI-fuelled election campaigns are here — where are the rules?, *Nature 628(8007)*, 237-237. ***https://www.nature.com/articles/d41586-024-00995-9***.

Ciancaglini, V., 2020, *Malicious Uses and Abuses of Artificial Intelligence*: Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI), and Europol's European Cybercrime Centre (EC3), ***https://www.europol.europa.eu/newsroom/news/new-report-finds-criminals-leverage-ai-for-malicious-use-%E2%80%93-and-it%E2%80%99s-not-just-deep-fakes***.

Coeckelbergh, M., 2024, *Why AI undermines Democracy and what to do about it*, Cambridge/Hoboken: Polity Press.

Coldewey, D., 2024, Amazon doubles down on Anthropic, completing its planned $4B investment, *TechCrunch*, March 27, ***https://techcrunch.com/2024/03/27/amazon-doubles-down-on-anthropic-completing-its-planned-4b-investment/?guccounter=1***.

Costa, A. F. und Coelho, N. M., 2024, Evolving Cybersecurity Challenges in the Age of AI-Powered Chatbots: A Comprehensive Review, in: Ferrada, F. und Camarinha-Matos, L. M. (Hg.): *Technological Innovation for Human-Centric Systems: 15th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2024, Caparica, Portugal, July 3–5, 2024, Proceedings*, Cham: Springer, 217-229.

Cotter, K., 2022, Selling Political Data: How Political Ad Tech Firms' Discourses Legitimate Microtargeting, in: Smits, M. (Hg.): *Information for a Better World: Shaping the Global Future*: Springer, 195-208, DOI:10.1007/978-3-030-96957-8_18.

Crowston, K. und Bolici, F., 2024, *Deskilling and upskilling with generative AI systems*, Working paper, ***https://crowston.syr.edu/node/1681***.

Cupać, J. und Sienknecht, M., 2024, Regulate against the machine: how the EU mitigates AI harm to democracy, *Democratization 31(5)*, 1067-1090. ***https://doi.org/10.1080/13510347.2024.2353706***.

Dale, R., 2021, GPT-3: What's it good for?, *Natural Language Engineering 27(1)*, 113-118. ***https://www.cambridge.org/core/journals/natural-language-engineering/article/gpt3-whats-it-good-for/0E05CFE68A7AC8BF794C8ECBE28AA990***.

Damer, N., Saladie, A. M., Braun, A. und Kuijper, A., 2018, MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network, *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1-10. ***https://ieeexplore.ieee.org/document/8698563/***.

Darius, P. und Römmele, A., 2023, KI und datengesteuerte Kampagnen: Eine Diskussion der Rolle generativer KI im politischen Wahlkampf: *Informationsflüsse, Wahlen und Demokratie*: Nomos Verlagsgesellschaft mbH & Co. KG, 199-212.

Davey, A., 2022, OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails, *Bloomberg.com*, 2022/12/08/, ***https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results***.

David, P. A., 1985, Clio and the Economics of QWERTY, *The American Economic Review 75(2)*, 332-337. ***http://www.jstor.org/stable/1805621***.

Debunk, 2023, *About Debunk: Disinformation Analysis Center*, ***https://www.debunk.org/about***.

Denkler, T., 2021, Der Tod von George Floyd in der Rekonstruktion, *Süddeutsche.de*. ***https://www.sueddeutsche.de/politik/george-floyd-tod-polizeigewalt-videos-rekonstruktion-1.4928047***.

Deutscher Ethikrat, 2023, *Mensch und Maschine – Herausforderungen durch künstliche Intelligenz*: Deutscher Ethikrat, ***https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf***.

Diakopoulos, N. und Johnson, D., 2019, Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections, Rochester, NY, ***https://papers.ssrn.com/abstract=3474183***.

Dobber, T., Metoui, N., Trilling, D., Helberger, N. und de Vreese, C., 2020, Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?, *The International Journal of Press/Politics 26(1)*. ***https://doi.org/10.1177/1940161220944364***.

Dommett, K., Kefford, G. und Kruschinski, S., 2023, *Data-Driven Campaigning and Political Parties: Five Advanced Democracies Compared*: Oxford University Press, **https://doi.org/10.1093/oso/9780197570227.001.0001**.

Dreyer, S., Stanciu, E., Potthast, K. C. und Schulz, W., 2021, *Desinformation: Risiken, Regulierungslücken und adäquate Gegenmaßnahmen*; Wissenschaftliches Gutachten, im Auftrag von: Landesanstalt für Medien NRW, **https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Desinformation/Leibnitz-Institut_LFMNRW_GutachtenDesinformation.pdf**.

Dubois, E. und Blank, G., 2018, The echo chamber is overstated: the moderating effect of political interest and diverse media, *Information, Communication & Society*, 729–745. doi:10.1080/1369118X.2018.1428656.

Dumbrava, C. (EPRS – Wissenschaftlicher Dienst des Europäischen Parlaments), 2021, *Die Hauptrisiken sozialer Medien für die Demokratie. Risiken durch Überwachung, Personalisierung, Desinformation, Moderation und Mikrotargeting*; Eingehende Analyse, Nr. PE 698.84, Dezember, Brüssel: Europäisches Parlament, **https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/698845/EPRS_IDA(2021)698845_DE.pdf**.

EGE (European Group on Ethics in Science and New Technologies), 2023, *Democracy in the Digital Age*; Opinion, im Auftrag von: DG Research and Innovation, Nr. 33, 20 June, Brussels: European Commission, doi:10.2777/078780.

EirGrid, S., 2012, All-island generation capacity statement 2013-2022', *EirGrid and System Operator of Northern Ireland (SONI)*. **https://cms.eirgrid.ie/sites/default/files/publications/All-Island_GCS_2013-2022.pdf**.

Eloundou, T., Manning, S., Mishkin, P. und Rock, D., 2024, GPTs are GPTs: Labor market impact potential of LLMs, *Science 384(6702)*, 1306-1308. **https://www.science.org/doi/10.1126/science.adj0998**.

EPTA (European Parliamentary Technlogy Assessment), 2023, *Generative Artificial Intelligence. Opportunities, Risks and Policy challenges. EPTA Report 2023*, Barcelona: CAPCIT, **https://www.parlament.cat/document/composicio/394503200.pdf**.

Europäische Kommission, 2018, Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen *Bekämpfung von Desinformation Im Internet: Ein Europäisches Konzept*, **https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52018DC0236&from=DE**.

Europäische Kommission, 2023, *Empfehlung vom 12.12.2023 zur Förderung der Mitwirkung und wirksamen Beteiligung von Bürgerinnen und Bürgern und Organisationen der Zivilgesellschaft an politischen Entscheidungsprozessen*, **https://commission.europa.eu/document/fcb629fe-ca20-4019-b1f6-392c286fdedf_en**.

Fallis, D., 2015, What Is Disinformation?, *Library Trends 63(3)*, 401–426. **https://doi.org/10.1353/lib.2015.0014**.

Farid, H. und Schindler, H.-J., 2020, *Die Gefahr von Deep Fakes für unsere Demokratie*, 29.06.: Konrad Adenauer Stiftung, **https://www.kas.de/de/einzeltitel/-/content/die-gefahr-von-deep-fakes-fuer-unsere-demokratie**.

Fletcher, R., Cornia, A., Graves, L. und Nielsen, R. K., 2018, *Measuring the reach of "fake news" and online disinformation in Europe*; Factsheet, im Auftrag von: Google, February: Reutters Institute and University of Oxford, **https://reutersinstitute.politics.ox.ac.uk/our-research/measuring-reach-fake-news-and-online-disinformation-europe**.

Fm1Today, 2023, Glarner lässt Arslan mit KI gegen «kriminelle Türken» hetzen, *FM1Today – regionale News aus der Ostschweiz*. **https://www.fm1today.ch/schweiz/glarner-laesst-arslan-mit-ki-gegen-kriminelle-tuerken-hetzen-154236739**.

Forster, K., 2003, Rezeption von Bildmanipulationen, in: Knieper, T. und Müller, M. G. (Hg.): *Authentizität und Inszenierung von Bilderwelten*, Köln: Herbert von Halem Verlag, 66-101.

FORWIT, 2024, *Empfehlungen für die FTI- und Wissenschaftspolitik einer Bundesregierung in der XXVIII. Legislaturperiode*, 13.09., Wien: FORWIT, **https://fti-monitor.forwit.at/docs/pdf/R000003.pdf**.

FORWIT und Beirat für Künstliche Intelligenz, 2024, *Empfehlung für die Schaffung von Rahmenbedingungen zur optimalen Entwicklung und Nutzung von Technologien der Künstlichen Intelligenz*, 08.10., **https://fti-monitor.forwit.at/docs/pdf/R000004.pdf**.

Gabriel, J. und Hadeed, M. (upgrade democracy), 2024, *Digitale Diskurse und demokratische Öffentlichkeit 2035*; Ergebnisbericht, 2024-07, Gütersloh: Bertelsmann Stiftung, **https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/digitale-diskurse-und-demokratische-oeffentlichkeit-in-2035-ergebnisbericht?tx_rsmbstpublications_pi2%5BfilterPreis%5D=0&cHash=7e6536c168dbd4300faf02057c9662d1**.

Godulla, A., Hoffmann, C. P. und Seibert, D., 2021, Dealing with deepfakes – an interdisciplinary examination of the state of research and implications for communication studies, *Studies in Communication and Media 10(1)*, 72-96. ***https://www.nomos-elibrary.de/index.php?doi=10.5771/2192-4007-2021-1-72***.

Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V. und Sikdar, B., 2024, Privacy and Security Concerns in Generative AI: A Comprehensive Survey, *IEEE Access 12*, 48126-48144. ***https://ieeexplore.ieee.org/document/10478883***.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. und Bengio, Y., 2014, Generative Adversarial Networks: arXiv, ***https://arxiv.org/abs/1406.2661***.

Grant, N., 2024, Google Fires 28 Employees Involved in Protest of Israeli Cloud Contract, *New York Times*, April 18, 2024, ***https://www.nytimes.com/2024/04/18/technology/google-firing-israeli-cloud-contract.html***.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. und Fritz, M., 2023, Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection: *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79-90, ***https://dl.acm.org/doi/10.1145/3605764.3623985***.

Groh, M., Epstein, Z., Firestone, C. und Picard, R., 2022, Deepfake detection by human crowds, machines, and machine-informed crowds, *Proceedings of the National Academy of Sciences 119(1)*, e2110013119. ***https://www.pnas.org/doi/10.1073/pnas.2110013119***.

Grünke, D. P., Litsche, S. und Starchenko, S., 2024, *Demokratiekompetenz stärken. Herausforderung Künstliche Intelligenz und die Vermittlung von Medienkompetenz*; Gutachten, 2024-03, Berlin: die medienanstalten, ***https://www.die-medienanstalten.de//fileadmin/user_upload/die_medienanstalten/Service/Studien_und_Gutachten/GVK_Gutachten_Demokratiekompetenz_st%C3%A4rken_2024.pdf***.

Gupta, M., Akiri, C., Aryal, K., Parker, E. und Praharaj, L., 2023, From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy, *IEEE Access 11*, 80218-80245. ***https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10198233***.

Gupta, N., 2020, Disadvantages of GANs || Am I real or a Trained Model to write?, *OpenGenus IQ: Computing Expertise & Legacy*, 18.11., ***https://iq.opengenus.org/disadvantages-of-gans/***.

Gupta, N., Ibañez, J. und Tenove, C., 2024, *The Peril and Promise of AI for Journalism*: University of Vancouver: Centre for the Study of Democratic Institutions,, ***https://ccss.arts.ubc.ca/research/report-the-peril-and-promise-of-ai-for-journalism/***.

Hackenburg, K. und Margetts, H., 2024, Evaluating the persuasive influence of political microtargeting with large language models, *Proceedings of the National Academy of Sciences 121(24)*, e2403116121. ***https://www.pnas.org/doi/abs/10.1073/pnas.2403116121***.

Haller, A. und Kruschinski, S., 2020, Politisches Microtargeting. Eine normative Analyse von datenbasierten Strategien gezielter Wähler_innenansprache, *Communicatio Socialis 53(4)*, 519-553. ***https://doi.org/10.5771/0010-3497-2020-4-519%20***.

Harari, Y. N., 2024, *Nexus. Eine kurze Geschichte der Informationsnetzwerke von der Steinzeit bis zur künstlichen Intelligenz*, München: Penguin.

Harwell, D., 2019, Faked Pelosi videos, slowed to make her appear drunk, spread across social media, *Washington Post*, 2019/05/24/, ***https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/***.

Hazell, J., 2023, *Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns*, arXiv:2305.06972 [cs.CY], ***https://arxiv.org/abs/2305.06972***.

Heap, B., Hansen, P. und Gill, M., 2021, *Strategic Communications Hybrid Threats Toolkit – Applying the principles of NATO Strategic Communications to understand and counter grey zone threats*: NATO Strategic Communications Centre of Excellence, ***https://books.google.at/books?id=YvpB0AEACAAJ***.

Heesen, J., Bieber, C., Grunwald, A., Matzner, T. und Roßnagel, A., 2021, *KI-Systeme und die individuelle Wahlentscheidung – Chancen und Herausforderungen für die Demokratie, Whitepaper aus der Plattform Lernende Systeme*; Whitepaper, September, München: Lernende Systeme – Die Plattform für Künstliche Intelligenz, ***https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_WP_KI_und_Wahlen.pdf***.

Hegelich, S. und Serrano, J. C. M., 2019, *Microtargeting in Deutschland bei der Europawahl 2019*; Bericht eines Kooperationsprojekts, Oktober, Düsseldorf: Landesanstalt für Medien NRW, ***https://www.medienanstalt-nrw.de/fileadmin/user_upload/lfm-nrw/Foerderung/Forschung/Dateien_Forschung/Studie_Microtargeting_DeutschlandEuropawahl2019_Hegelich_web2.pdf***.

Heil, R., 2023, *Einige ethische Implikationen großer Sprachmodelle*, KIT Scientific Working Paper Nr. 221: Karlsruher Institut für Technologie, ***https://publikationen.bibliothek.kit.edu/1000158914***.

High-Level Expert Group on AI, 2019, *Ethics Guidelines for Trustworthy AI*, 8 April 2019, Brussels: European Commission, ***https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1***.

HND BW (Hochschulnetzwerk Digitalisierung der Lehre Baden-Württemberg), 2024, *KI in der Hochschullehre, 11. März 2024, DHBW Stuttgart*, ***https://www.hnd-bw.de/veranstaltungen/ki-lehre/***.

Holland, M., 2023, ChatGPT & Co: Dutzende Seiten mit KI-generierten nachrichtentexten entdeckt, *Heise online*, 02.05., ***https://www.heise.de/news/ChatGPT-Co-Dutzende-Seiten-mit-generierten-nachrichtentexten-entdeckt-8984184.html***.

Holzer, S. und Sengl, M., 2020, Quelle gut, alles gut? Glaubwürdigkeitsbeurteilung im digitalen Raum, in: Hohlfeld, R., Harnischmacher, M., Heinke, E., Lehner, L. und Sengl, M. (Hg.): *Fake News und Desinformation: Herausforderungen für die vernetzte Gesellschaft und die empirische Forschung*, Baden-Baden: Nomos, 155-178, ***https://www.nomos-elibrary.de/10.5771/9783748901334-155/quelle-gut-alles-gut-glaubwuerdigkeitsbeurteilung-im-digitalen-raum?page=1***.

Horn, M. B., 2024, Artificial Intelligence, Real Anxiety, *Education Next 24(2)*, 1-1. ***https://www.educationnext.org/artificial-intelligence-real-anxiety-how-should-educators-use-ai-prepare-students-future/***.

Howard, P. N., 2020, *Lie Machines*, Yale: Yale University Press.

Hügel, S., 2019, Künstliche Intelligenz und Politik, *Vorgänge 225/226(1.2)*, 25-42. ***https://www.humanistische-union.de/publikationen/vorgaenge/225-226/publikation/kuenstliche-intelligenz-und-politik/***.

HuggingFace, 2024, *Models*; [Aufgerufen am: 15.05. 2024], ***https://huggingface.co/models?p=1&sort=trending***.

Hurz, S., 2021, CDU blamiert sich mit Anzeige gegen IT-Expertin, *Süddeutsche Zeitung*, ***https://www.sueddeutsche.de/politik/cdu-connect-anzeige-wittmann-1.5373488***.

International Telecommunication Union, 2020, *Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris agreement*, IT-U Reccomendation Nr. L. 1470. (01/20), ***http://handle.itu.int/11.1002/1000/14084***.

Internationale Hochschule (IU), 2023, Studie: Ein Drittel der Arbeitenden erwarten sich durch ChatGPT & Co. Erleichterung im Joballtag, ***https://www.iu.de/news/studie-auswirkungen-von-chatgpt-und-ki-bots-auf-die-arbeitswelt/***.

Jäger, W., Nentwich, M., Embacher-Köhle, G. und Krieger-Lamina, J., 2022, Kann es eine digitale Souveränität Österreichs geben? Herausforderungen für den Staat in Zeiten der Digitalen Transformation, in: Bogner, A., Decker, M., Nentwich, M. und Scherz, C. (Hg.): *Digitalisierung und die Zukunft der Demokratie. Beiträge aus der Technikfolgenabschätzung*, Berlin: Nomos, 189-204, ***https://www.nomos-elibrary.de/10.5771/9783748928928-189.pdf?download_chapter_pdf=1&page=1***.

Janeway, W. H., 2012, *Doing capitalism in the innovation economy: Markets, speculation and the state*, Cambridge: CUP.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Dai, W., Madotto, A. und Fung, P., 2023, Survey of Hallucination in Natural Language Generation, *ACM Computing Surveys 55(12)*, 1-38. ***http://arxiv.org/abs/2202.03629***.

Jonas, U. und Marinov, V., 2022, Deepfakes zwischen Pornografie und politischer Desinformation – wie groß ist die Gefahr?, *correctiv.org*, ***https://correctiv.org/faktencheck/hintergrund/2022/08/12/deepfakes-zwischen-pornografie-giffey-und-politischer-desinformation-wie-gross-ist-die-gefahr/***.

Jumaev, G., 2024, The Impact of AI on Job Market: Adapting to the Future of Work, *Modern Science and Research 3(1)*. ***https://inlibrary.uz/index.php/science-research/article/view/28146***.

Jungherr, A., 2023, Digital campaigning: how digital media change the work of parties and campaign organizations and impact elections, in: Skopek, J. (Hg.): *Research Handbook on Digital Sociology*: Edward Elgar Publishing, 446-462, ***https://www.elgaronline.com/view/book/9781789906769/book-part-9781789906769-35.xml***.

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F. und Rolnick, D., 2022, Aligning artificial intelligence with climate change mitigation, *Nature Climate Change 12(6)*, 518-527. ***https://www.nature.com/articles/s41558-022-01377-7***.

Kapantai, E., Christopoulou, A., Berberidis, C. und Peristeras, V., 2021, A Systematic Literature Review on Disinformation: Toward a Unified Taxonomical Framework, *New Media & Society 23 (5)*, 1301-1326. ***https://journals.sagepub.com/doi/abs/10.1177/1461444820959296***.

Karaboga, M., 2023, Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags, in: Jaki, S. und Steiger, S. (Hg.): *Digitale Hate Speech: In-terdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation*: Springer, 197–220, ***https://doi.org/10.1007/978-3-662-65964-9_10***.

Karaboga, M., Frei, N., Puppis, M., Vogler, D., Raemy, P., Ebbers, F., Runge, G., Rauchfleisch, A., de Seta, G., Gurr, G., Friedewald, M. und Rovelli, S., 2024, *Deepfakes und manipulierte Realitäten. Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz*; in Reihe: TA-SWISS Publikationsreihe, Bd. TA 81/2024, hg. v. TA-Swiss, Zollikon: vdf, ***https://vdf.ch/deepfakes-und-manipulierte-realitaten-e-book.html***.

Karmasin, M., Pöschl, M., Prainsack, B., Puntscher-Riekmann, S. und Strauß, S. (Österreichische Akademie der Wissenschaften), 2024, *Sind Soziale Medien eine Gefahr für unsere Demokratie?*; Stellungnahme der Ad-hoc-Arbeitsgruppe, Wien: ÖAW, ***https://www.oeaw.ac.at/fileadmin/NEWS/2024/pdf/aid-fug-6-2024.pdf***.

Khanal, S., Zhang, H. und Taeihagh, A., 2024, Why and how is the power of Big Tech increasing in the policy process? The case of generative AI, *Policy and Society*, puae012. ***https://doi.org/10.1093/polsoc/puae012***.

Kietzmann, J., Lee, L. W., McCarthy, I. P. und Kietzmann, T. C., 2020, Deepfakes: Trick or treat?, *Business Horizons 63(2)*, 135-146. ***https://www.sciencedirect.com/science/article/pii/S0007681319301600***.

Kollapally, N. M. und Geller, J., 2024, Safeguarding Ethical AI: Detecting Potentially Sensitive Data Re-Identification and Generation of Misleading or Abusive Content from Quantized Large Language Models, *BIOSTEC (2)*, ***https://www.scitepress.org/publishedPapers/2024/124119/pdf/index.html***.

Kopcik, P. M., 2023, *Politisches Microtargeting auf Social Media: Regulierung von digitalen Wahlkampagnen in Österreich aus datenschutzrechtlicher Perspektive*, Masterarbeit, Internationale Wirtschaftsbeziehungen, FH Burgenland Eisenstadt, ***https://fhburgenland.contentdm.oclc.org/digital/api/collection/p15425dc/id/149883/download***.

Krieger-Lamina, J. und Peissl, W., 2024, im Erscheinen, *Privatsphäre 3.0. Konsument:innen im digitalen Raum*; Endbericht, im Auftrag von: Arbeiterkammer Wien, Nr. ITA-2024-04, Mai, Wien: Institut für Technikfolgen-Abschätzung.

Kruschinski, S., 2023, Wie uns Parteien mit Daten und Technologien manipulieren, *Colloquium Fundamentale mit dem Thema "Polarisiert und desinformiert? Politische Information im digitalen Zeitalter*, 14.12., KIT, Karlsruhe, ***https://www.youtube.com/watch?v=cat6SHlrqlc***.

Łabuz, M., 2024, Deep fakes and the Artificial Intelligence Act—An important signal or a missed opportunity?, *Policy & Internet n/a(n/a)*, 1-18. ***https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.406***.

Lallie, H. S., Shepherd, L. A., Nurse, J. R. C., Erola, A., Epiphaniou, G., Maple, C. und Bellekens, X., 2021, Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic, *Computers & Security 105*, 102248. ***https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9755115/***.

Lasoen, K., 2022, *Realising the EU Hybrid Toolbox: opportunities and pitfalls*, Clingendael Policy Briefs: Clingendael Institute – the Netherlands Institute of International Relations, ***https://www.clingendael.org/sites/default/files/2022-12/Policy_brief_EU_Hybrid_Toolbox.pdf***.

Lee, K., Cooper, A. F. und Grimmelmann, J., 2023, *Talkin''Bout AI Generation: Copyright and the Generative-AI Supply Chain*, arXiv preprint arXiv:2309.08133, ***https://www.law.berkeley.edu/wp-content/uploads/2024/01/Talkin-Bout-AI-Generation-1.pdf***.

Leisegang, D., 2023, *Prekäre Klickarbeit hinter den Kulissen von ChatGPT*; 2024], ***https://netzpolitik.org/2023/globaler-sueden-prekaere-klickarbeit-hinter-den-kulissen-von-chatgpt/***.

Lemley, M., 2024, How Generative AI Turns Copyright Law Upside Down, *Science and Technology Law Review 25(2)*. ***https://doi.org/10.52214/stlr.v25i2.12761***.

Li, L., Bao, J., Yang, H., Chen, D. und Wen, F., 2020, FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping: arXiv, ***http://arxiv.org/abs/1912.13457***.

Li, P., Yang, J., Islam, M. A. und Ren, S., 2023, Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models, *arXiv preprint arXiv:2304.03271*. **https://arxiv.org/abs/2304.03271**.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. und Zou, J., 2023, *GPT detectors are biased against non-native English writer*, arXiv, **https://arxiv.org/pdf/2304.02819**.

Libiseller, C., 2023, 'Hybrid warfare' as an academic fashion, *Journal of Strategic Studies 46(4)*, 858-880. **https://doi.org/10.1080/01402390.2023.2177987**.

Lin, B. und Lewis, S. C., 2022, The one thing journalistic AI just might do for democracy, *Digital journalism 10(10)*, 1627-1649. **https://doi.org/10.1080/21670811.2022.2084131**.

López Ortega, A., 2022, Are microtargeted campaign messages more negative and diverse? An analysis of Facebook Ads in European election campaigns, *European Political Science 21(3)*, 335-358. **https://doi.org/10.1057/s41304-021-00346-6**.

Luccioni, A. S., Viguier, S. und Ligozat, A.-L., 2023, Estimating the carbon footprint of bloom, a 176b parameter language model, *Journal of Machine Learning Research 24(253)*, 1-15. **https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf**.

Madeira, O., Hansen, J.-H. und Kolleck, A. (TAB – Büro für Technikfolgenabschätzung beim Deutschen Bundestag), 2024, *Deepfakes – legal and societal challenges as well as innovation potentials*; Report, Berlin: Deutscher Bundestag,, **https://www.tab-beim-bundestag.de/english/projects_deepfakes-legal-and-societal-challenges-as-well-as-innovation-potentials.php**.

Maguire, J. und Ross Winthereik, B., 2021, Digitalizing the state: Data centres and the power of exchange, *Ethnos 86(3)*, 530-551. **https://doi.org/10.1080/00141844.2019.1660391**.

Maier, T. und Schmid, F., 2022, Video von Ludwigs Telefonat mit falschem Klitschko ging online, *DER STANDARD*. **https://www.derstandard.de/story/2000138143911/video-von-ludwigs-treffen-mit-falschem-klitschko-ging-online**.

Malmodin, J. und Lundén, D., 2018, The energy and carbon footprint of the global ICT and E&M sectors 2010–2015, *Sustainability 10(9)*, 3027.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M. und Sarkar, R., 2024, Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet, in: Cuzzolin, F. und Sultana, M. (Hg.): *Epistemic Uncertainty in Artificial Intelligence*, Cham: Springer, **https://doi.org/10.1007/978-3-031-57963-9_5**.

Matasick, C., Villanova, N., Zdanavicius, L. und Baubion, C. (OECD Directorate for Public Governance (GOV) – Anti-Corruption and Integrity in Government Division), 2024, *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity* im Auftrag von: Organisation for Economic Co-operation and Development, March, Paris: OECD, **https://doi.org/10.1787/d909ff7a-en**.

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M. und Cerf, M., 2024, The potential of generative AI for personalized persuasion at scale, *Scientific Reports 14(1)*, 4692. **https://www.nature.com/articles/s41598-024-53755-0**.

Mazzucato, M., 2018, Mission-oriented innovation policies: challenges and opportunities, *Industrial and corporate change 27(5)*, 803-815.

Mazzucchi, N., 2022, *AI-based technologies in hybrid conflict: The future of influence operations*, Hybrid CoE Paper Nr. 14, **https://www.hybridcoe.fi/publications/hybrid-coe-paper-14-ai-based-technologies-in-hybrid-conflict-the-future-of-influence-operations/**.

Merli, F., 2019, Grenzen der Staatsinformation und staatlicher Propaganda, in: Berka, Holoubek und Leitl-Staudinger (Hg.): *Elektronische Medien im „postfaktischen" Zeitalter – Dreizehntes Rundfunkforum*, Wien Manz, 107-120.

Microsoft, 2024, *2024 Environmental Sustainability Report*, **https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lMjE**.

Milanez, A., 2023, The impact of AI on the workplace: Evidence from OECD case studies of AI implementation. **https://www.oecd-ilibrary.org/content/paper/2247ce58-en**.

Milmo, D., 2024, Google's emissions climb nearly 50% in five years due to AI energy demand, *The Guardian*, **https://www.theguardian.com/technology/article/2024/jul/02/google-ai-emissions**.

Miró-Llinares, F. und Aguerri, J. C., 2023, Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat', *European Journal of Criminology 20(1)*, 356-374. **https://journals.sagepub.com/doi/abs/10.1177/1477370821994059**.

MIT CSAIL, 2017, *Wikum: Bridging Discussion Systems and Wikis with Collective Summarization*, ***https://www.csail.mit.edu/research/wikum-bridging-discussion-systems-and-wikis-collective-summarization***.

Monopol, 2023, Bundesregierung geht offenbar gegen Deepfake-Video von Olaf Scholz vor, *Monopol*, 30.11., ***https://www.monopol-magazin.de/bundesregierung-geht-offenbar-gegen-deepfake-video-von-olaf-scholz-vor***.

Muldoon, J., Cant, C., Graham, M. und Ustek Spilda, F., 2023, The poverty of ethical AI: impact sourcing and AI supply chains, *AI & Society*. ***https://doi.org/10.1007/s00146-023-01824-9***.

Müller-Brehm, J., 2019, *Forschungsstand: Microtargeting in Deutschland und Europa – Fehlende Transparenz und viele offene Fragen*; Bericht, Düsseldorf: Landesanstalt für Medien NRW, ***https://www.medienanstalt-nrw.de/fileadmin/user_upload/lfm-nrw/Foerderung/Forschung/Dateien_Forschung/Forschungsmonitoring_Microtargeting_Deutschland_Europa.pdf***.

Murray, M. D., 2023, Generative ai art: Copyright infringement and fair use, *SMU Science and Technology Law Review 26(2)*, 259. ***https://doi.org/10.25172/smustlr.26.2.4***.

Narechania, T. N., 2021, Machine Learning as Natural Monopoly, *Iowa L. Rev. 107*, 1543.

Nentwich, M., Jäger, W., Embacher-Köhle, G. und Krieger-Lamina, J., 2019, *Kann es eine digitale Souveränität Österreichs geben? Herausforderungen für den Staat in Zeiten der Digitalen Transformation*, ITA Manu:scripts Nr. ITA-19-01, ***http://epub.oeaw.ac.at/ita/ita-manuscript/ita_19_01.pdf***.

Newman, N. (Reuters Institute), 2023, *Overview and key findings of the 2023 Digital News Report*, ***https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/dnr-executive-summary***.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V. und Nguyen, C. M., 2022, Deep Learning for Deepfakes Creation and Detection: A Survey, *Computer Vision and Image Understanding 223*, 103525. ***https://doi.org/10.1016/j.cviu.2022.103525***.

Nicolaus, K., 2023, Dieses Bild von einem Mann mit fünf Kindern in Gaza wurde mit Künstlicher Intelligenz generiert, *correctiv.org*, ***https://correctiv.org/faktencheck/2023/11/08/dieses-bild-von-einem-mann-mit-fuenf-kindern-in-gaza-wurde-mit-kuenstlicher-intelligenz-generiert/***.

O'Brien, T. C. und Tyler, T. R., 2020, Authorities and communities: Can authorities shape cooperation with communities on a group level?, *Psychology, Public Policy, and Law 26(1)*, 69.

Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G. L. und Wang, Y., 2023, Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare, *NPJ Digital Medicine 6(1)*, 225.

Oremus, W., 2023, AI chatbots lose money every time you use them. That is a problem., *The Washington Post*, June 5, 2023, ***https://www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/***.

PA Media, 2022, *RADAR: Combining the latest in AI with skilled writers to dynamically create high-quality content at massive scale*, ***https://pa.media/radar/***.

Pariser, E., 2012, *Filter Bubble. Wie wir im Internet entmündigt werden*, München Hanser.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. und Chintala, S., 2019, PyTorch: An Imperative Style, High-Performance Deep Learning Library: arXiv, ***http://arxiv.org/abs/1912.01703***.

Patel, D., 2023, *The AI Brick Wall – A Practical Limit For Scaling Dense Transformer Models, and How GPT 4 Will Break Past It*; Letzte Aktualisierung: Jan 24, 2023, ***https://www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit***.

Paulitsch, L., 2024, Gegenerzählungen für »Selberdenker« Ein Versuch der Einordnung von »Alternativmedien« im konservativen Spektrum, *Journalistik*, 2, ***https://journalistik.online/ausgabe-2024/gegenerzaehlungen-fuer-selberdenker/***.

Pawelec, M., 2022, Deepfakes als Chance für die Demokratie?, in: Bogner, A., Decker, M., Nentwich, M. und Scherz, C. (Hg.): *Digitalisierung und die Zukunft der Demokratie*: Nomos Verlagsgesellschaft mbH & Co. KG, 89-102, ***https://www.nomos-elibrary.de/index.php?doi=10.5771/9783748928928-89***.

Pig, C., 2023, *Democracy Dies in Darkness*, Wien: Christian Brandstätter Verlag.

Posetti, J., 2018, Combatting online abuse: When journalists and their sources are targeted, in: Cherilyn, I. und Julie, P. (Hg.): *Journalism, fake news & disinformation: Handbook for journalism education and training* Paris: UNESCO Publishing, 115–127.

Posner, R. A., 1978, Natural monopoly and its regulation, *J. Reprints Antitrust L. & Econ. 9*, 767. ***https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=2883&context=journal_articles***.

Prummer, A., 2020, Micro-targeting and polarization, *Journal of Public Economics 188(August)*, 104210 ***https://sciencedirect.com/science/article/pii/S0047272720300748***.

Queck, S. und Oppelt, J., 2018, Microtargeting – Definition, Einsatz und Beispiele *marconomy*, 06.08., ***https://www.marconomy.de/microtargeting-definition-einsatz-und-beispiele-a-739666/***.

Quintais, J. P., 2024, *Generative AI, copyright and the AI Act, v.2*, SSRN, ***https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4912701***.

Rauschenberger, S., Ali-Pahlavani, N., Andreasch, D., Ernest, S., Franek, J., Jäger, A. und Schmatz, P., 2023, *Der Medienkompetenz-Bericht 2023*, 2023, Wien: Rundfunk und Telekom Regulierungs-GmbH, ***https://www.rtr.at/medien/aktuelles/publikationen/Publikationen/RTR_Medienkompetenzbericht_2023.pdf***.

Reveland, C. und Siggelkow, P., 2023, Künstliche Intelligenz: Falsche Tagesschau-Audiodateien im Umlauf, *tagesschau.de*, ***https://www.tagesschau.de/faktenfinder/tagesschau-audio-fakes-100.html***.

Rodel, L., Hertog, E., Unruh, C. und Susskind, D., 2024, AI and the Future of Work Workshop Report, *Institute for Ethics in AI, University of Oxford* 4. ***https://www.oxford-aiethics.ox.ac.uk/sites/default/files/2024-01/Report%20-%20AI%20and%20the%20Future%20of%20Work%20Workshop%20-%204%20and%205%20December%202023.pdf***.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. und Lewandowsky, S., 2022, Psychological inoculation improves resilience against misinformation on social media, *Science Advances 8(34)*, eabo6254. ***https://www.science.org/doi/abs/10.1126/sciadv.abo6254***.

Rozado, D., 2023, The Political Biases of ChatGPT, *Social Sciences 12(3)*, 148. ***https://www.mdpi.com/2076-0760/12/3/148***.

Rudnicki, K. und Steiger, S., 2020, *Online hate speech: Introduction into motivational causes, effects and regulatory contexts*, Antwerpen: Media Diversity Institute, ***https://www.media-diversity.org/wp-content/uploads/2020/09/DeTact_Online-Hate-Speech.pdf***.

Samuelson, P., 2023, Generative AI meets copyright. Ongoing lawsuits could affect everyone who uses generative AI, *Science 381*, 6654. ***http://dx.doi.org/10.1126/science.adi0656***.

Saurwein, F., Spencer-Smith, C. und Krieger-Lamina, J., 2022, Social-Media-Algorithmen als Gefahr für Öffentlichkeit und Demokratie: Anwendungen, Risikoassemblagen und Verantwortungszuschreibungen, in: Bogner, A., Decker, M., Nentwich, M. und Scherz, C. (Hg.), Baden-Baden: Nomos, 243-256, ***https://www.nomos-elibrary.de/10.5771/9783748928928-243.pdf?download_chapter_pdf=1&page=1***.

Saxer, U., 2023, *Von den Medien zu den Plattformen. Die Regulierung öffentlicher Kommunikation im Zeichen der digitalen Revolution* in Reihe: Schriften zum Medienrecht und Kommunikationsrecht (SMKR), Bd. 12, Tübingen: Mohr-Siebeck, ***https://www.mohrsiebeck.com/buch/von-den-medien-zu-den-plattformen-9783161623240/***.

Schell, K., 2022, *Journalistische Textautomatisierung – Status, Potenziale, Limitationen* im Auftrag von: APA – Austria Presse Agentur, Vienna, ***https://apa.at/service/whitepaper/***.

Schiller, D., 1999, *Digital capitalism: Networking the global market system*: MIT press.

Schiller, D., 2014, Rosa Luxemburg's Internet? For a Political Economy of State Mobilization and the Movement of Accumulation in Cyberspace, *International Journal of Communication (19328036) 8*.

Schumpeter, J. A., 1942[2013], *Capitalism, socialism and democracy*, 3. Aufl.: Routledge.

Schünemann, W. J., 2022, A threat to democracies?: An overview of theoretical approaches and empirical measurements for studying the effects of disinformation, in: Dunn Cavelty, M. und Wenger, A. (Hg.): *Cyber Security Politics*: Routledge, 32-47, ***https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003110224-4/threat-democracies-wolf-sch%C3%BCnemann***.

Sharir, O., Peleg, B. und Shoham, Y., 2020, The Cost of Training NLP Models: A Concise Overview: arXiv, ***http://arxiv.org/abs/2004.08900***.

Sharma, A., Sharma, A., Yaduvanshi, E. und Bhowal, I., 2024, In Machines We Trust: Anthropomorphism's Role in the Subtle Erosion of Human Expertise, *Journal of Computational Analysis and Applications (JoCAAA) 33(2)*, 380-384.

Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J., Masanet, E., Horner, N., Azevedo, I. und Lintner, W., 2016, *United states data center energy usage report*, Nr. LBNL-1005775, June: Ernesto Orlando Lawrence Berkley National Laboratory, ***https://www.iea-4e.org/wp-content/uploads/publications/2016/06/05j_-_LBNL_-_US_Data_Centres_Energy_USe.pdf***.

Shivakumar, S., Wessner, C. und Howell, T., 2024, *Balancing the Ledger – Export Controls on U.S. Chip Technology to China*, 21.02.: Center for Strategic and International Studies, ***https://www.csis.org/analysis/balancing-ledger-export-controls-us-chip-technology-china***.

Siddik, M. A. B., Shehabi, A. und Marston, L., 2021, The environmental footprint of data centers in the United States, *Environmental Research Letters 16(6)*, 064017.

Simchon, A., Edwards, M. und Lewandowsky, S., 2024, The persuasive effects of political microtargeting in the age of generative artificial intelligence, *PNAS Nexus 3(2)*. ***https://doi.org/10.1093/pnasnexus/pgae035***.

Simon, F. M., 2024, *Artificial Intelligence in the News: How AI Retools, Rationalizes, and Reshapes Journalism and the Public Arena*, 06.02.: Columbia Journalism School, ***https://towcenter.columbia.edu/sites/default/files/content/Tow%20Report_Felix-Simon-AI-in-the-News.pdf***.

Singh, T., 2023, *The impact of large language multi-modal models on the future of job market*, arXiv preprint arXiv:2304.06123, ***https://arxiv.org/abs/2304.06123***.

Spiegel Ausland, 2024, Sardinien: Italiens Regierungschefin Giorgia Meloni klagt gegen Porno-Fakes mit ihrem Gesicht, *Der Spiegel online*, 3.3., ***https://www.spiegel.de/ausland/sardinien-italiens-regierungschefin-giorgia-meloni-klagt-gegen-porno-fakes-mit-ihrem-gesicht-a-24a52211-1a68-4e89-a779-9a6877202d8c***.

STAA (Science Technology Assessment and Analytics), 2023, *Science & Tech Spotlight: Generative AI*; Policy Brief, Nr. GAO-23-106782, June, Washington D.C.: GAO, ***https://www.gao.gov/assets/gao-23-106782.pdf***.

Staab, P., 2019, *Digitaler Kapitalismus: Markt und Herrschaft in der Ökonomie der Unknappheit*: Suhrkamp Verlag.

Stanford Deliberative Democracy Lab, 2022, *Online Deliberation Platform*, ***https://deliberation.stanford.edu/tools-and-resources/online-deliberation-platform***.

Stöcker, C., 2024, Lernende Maschinen und die Zukunft der Öffentlichkeit, in: Schreiber, G. und Ohly, L. (Hg.): *KI:Text. Diskurse über KI-Textgeneratoren*, Berlin/Boston: De Gruyter, 401-418, ***https://doi.org/10.1515/9783111351490***.

Stokel-Walker, C. und Van Noorden, R., 2023, What ChatGPT and generative AI mean for science, *Nature 614(7947)*, 214-216. ***https://www.nature.com/articles/d41586-023-00340-6***.

Strauß, S., 2020, Vom „Global Village" zur „Blackbox Society"? Digitale Identitäten und politische Kommunikation in Zeiten des Überwachungskapitalismus, *Momentum Quarterly 9(9)*, 85-102. ***https://momentum-quarterly.org/momentum/article/view/3387/2677***.

Strauß, S. und Udrea, T., 2024, *CAIL – Critical AI Literacy (Endbericht)*, Wien: Institut für Technikfolgen-Abschätzung, ***https://epub.oeaw.ac.at/ita/ita-projektberichte/ITA-2024-07.pdf***.

Taecharungroj, V., 2023, "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter, *Big Data and Cognitive Computing 7(1)*, 35.

Tagesschau, 2023, Kehrt ChatGPT unter Bedingungen zurück?, ***https://www.tagesschau.de/ausland/europa/italien-rueckkehr-chatgpt-bedingungen-101.html***.

Tandoc, E. C., Lim, Z. W. und Ling, R., 2018, Defining "Fake News", *Digital Journalism (6)*, 137-147.

The Computational Democracy Project, 2024a, *Case-Studies: The Klimarat in Austria*, ***https://compdemocracy.org/Case-studies/2022-Austria-Klimarat/***.

The Computational Democracy Project, 2024b, *Pol.is*, ***https://compdemocracy.org/Polis/***.

Thiel, T., 2020, Demokratie in der digitalen Konstellation, in: Riescher, G., Rosenzweig, B. und Meine, A. (Hg.): *Einführung in die Politische Theorie: Grundlagen–Methoden–Debatten*, Stuttgart, 331-349.

Tiwari, R., 2023, The impact of AI and machine learning on job displacement and employment opportunities, *International Journal of Engineering Technologies and Management Research 7(1)*, 507-512. ***https://ijetms.in/Vol-7-issue-4/Vol-7-Issue-4-67.pdf***.

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C., Ezer, D., Haert, F. C. v. d. und Mugisha, F., 2020, AI for social good: unlocking the opportunity for positive impact, *Nature Communications 11(1)*, 2468. ***https://pubmed.ncbi.nlm.nih.gov/32424119/***.

Tsai, L. L., Morse, B. S. und Blair, R. A., 2020, Building credibility and cooperation in low-trust settings: persuasion and source accountability in Liberia during the 2014–2015 Ebola crisis, *Comparative Political Studies 53(10-11)*, 1582-1618.

Tsai, L. L., Pentland, A., Braley, A., Chen, N., Enríquez, J. R. und Reuel, A., 2024, Generative AI for Pro-Democracy Platforms. ***https://mit-genai.pubpub.org/pub/mn45hexw/release/1***.

U.S. Central Intelligence Agency, 2020, The world fact book — total water withdrawal. ***https://www.cia.gov/the-world-factbook/field/total-water-withdrawal/***.

Udrea, T., Fuchs, D. und Peissl, W., 2022, *Künstliche Intelligenz. Verstehbarkeit und Transparenz*; Endbericht, Wien: Institut für Technikfolgen-Abschätzung, ***https://epub.oeaw.ac.at/ita/ita-projektberichte/ITA-pb-2022-01.pdf***.

UNESCO, 2022, *Recommendation on the Ethics of Artificial Intelligence – UNESCO Digital Library*, ***https://unesdoc.unesco.org/ark:/48223/pf0000381137***.

Unzicker, K., 2023, *Desinformation: Herausforderung für die Demokratie. Einstellungen und Wahrnehmungen in Europa*, 10.08: Bertelsmann Stiftung, ***https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/desinformation-herausforderung-fuer-die-demokratie***.

Upgrade Democracy, 2023, *Glossar: Methoden zum Umgang mit digitaler Desinformation*; Letzte Aktualisierung: 2023-08-04, ***https://www.bertelsmann-stiftung.de/de/unsere-projekte/upgrade-democracy/projektnachrichten/glossar-methoden-zum-umgang-mit-digitaler-desinformation***.

Ustek Spilda, F., Brittain, L., Cant, C. und Graham, M., 2024, *The Unmagical World of AI: Workers at the bottom of the AI supply chain*, 26.02., Brussels: Friedrich Ebert Stiftung, ***https://futureofwork.fes.de/news-list/e/ai-value-chain.html***.

van Huijstee, M., van Boheemen, P., Das, D., Nierling, L., Jahnel, J., Karaboga, M. und Fatun, M., 2021, *Tackling deepfakes in European policy*, im Auftrag von: European Parliamentary Research Service & Scientific Foresight Unit: European Parliament, ***http://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf***.

Vasse'i, R. M. und Udoh, G., 2024, Watermarking, Content Labeling Struggle to Effectively Distinguish AI-Generated Content, *Mozilla Foundation*. ***https://foundation.mozilla.org/en/blog/mozilla-research-watermarking-content-labeling-struggle-to-effectively-distinguish-ai-generated-content/***.

Villar García, J. P., Tarín Quirós, C., Blázquez Soria, J., Galán Pascual, C. und Galá, C., 2021, *Strategic communications as a key factor in countering hybrid threats*; Study report, im Auftrag von: STOA, Nr. PE 656.323, March, Brussels: European Parliament, EPRS – European Parliamentary Research Service, ***https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)656323***.

Wang, T., Zhang, Y., Qi, S., Zhao, R., Xia, Z. und Weng, J., 2023, Security and Privacy on Generative Data in AIGC: A Survey: arXiv, ***http://arxiv.org/abs/2309.09435***.

Wang, W., Gao, G. und Agarwal, R., 2024, Friend or foe? Teaming between artificial intelligence and workers with variation in experience, *Management Science 70(9)*, 5753-5775. ***http://dx.doi.org/10.1287/mnsc.2021.00588***.

Wardle, C., 2019, *First Draft's essential guide to understanding information disorder*, First Draft, ***https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x76701***.

Wardle, C. und Derakhshan, H., 2017, *Information Disorder: Toward an Interdisciplinary Frame-work for Research and Policy Making*, im Auftrag von: Council of Europe, Nr. DGI(2017)09, ***https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c***.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O. und Waddington, L., 2023, Testing of detection tools for AI-generated text, *International Journal for Educational Integrity 19(1)*, 26. ***https://doi.org/10.48550/arXiv.2306.15666***.

WEF (World Economic Forum), 2024, *The Global Risks Report 2024. 19th Edition.*, ***https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf***.

Widder, D. G., West, S. und Whittaker, M., 2023, *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI*, SSRN, Rochester, NY, ***https://papers.ssrn.com/abstract=4543807***.

Wolfram, 2023, What Is ChatGPT Doing … and Why Does It Work? *https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/*.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. und Hazelwood, K., 2021, *Sustainable AI: Environmental Implications, Challenges and Opportunities*, arXiv, *https://arxiv.org/abs/2111.00364*.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. und Zhang, Y., 2024, A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly, *High-Confidence Computing 4(2)*. *https://www.sciencedirect.com/science/article/pii/S266729522400014X*.

Zarouali, B., Dobber, T., De Pauw, G. und de Vreese, C., 2022, Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media, *Communication Research 49(8)*, 1066-1091. *https://journals.sagepub.com/doi/abs/10.1177/0093650220961965*.

Zhang, H., Edelman, B. L., Francati, D., Venturi, D., Ateniese, G. und Barak, B., 2024, *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models*, arXiv, *https://arxiv.org/pdf/2311.04378*.

Zuboff, S., 2019, *The Age of Surveillance Capitalism. The Fight for the Future at the New Frontier of Power*, London: Profile Books.

WWW.OEAW.AC.AT