Jahrbuch [jtphil.nomos.de] Technikphilosophie 2025

Alpsancar | Friedrich | Gehring | Kaminski | Nordmann [Hrsg.]

Täuschung und Illusion

11. Jahrgang 2025

Nomos

Jahrbuch Technikphilosophie
11. Jahrgang 2025

Suzana Alpsancar | Alexander Friedrich
Petra Gehring | Andreas Kaminski
Alfred Nordmann [Hrsg.]

Täuschung und Illusion

Nomos

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de

ISBN 978-3-7560-3462-8 (Print)
978-3-7489-6497-1 (ePDF)

ISSN 2297-2072 (print)
2297-2080 (eBooks)

Redaktion / Editorial Team: Kai Denker, Dawid Kasprowicz,
Hildrun Lampe, Viet Anh Nguyen Duc, Tom Poljanšek, Željko Radinković, Jörn Wiengarn
Korrektorat / Copy Editors: Melina Licht, Arthur Wei-Kang Liu, Thorid Charlotte Schäfer, Lisa Schwarz

1. Auflage 2026

Inhaltsverzeichnis

Archiv

Diskussion

Glosse

Editorial

Keine guten Täuschungen oder Illusionen ohne gute Techniken. Jeder Trick braucht gutes Handwerk, jedes Blendwerk die passende Apparatur und die richtigen Werkzeuge: die Retusche, die Bühnenkulisse, den Algorithmus oder geschickte rhetorische Handgriffe. Magierinnen, Kunstfälscher, Computerspieldesignerinnen und Propagandisten wissen, dass technisches Können und die richtigen technischen Mittel den Unterschied zwischen billiger Kopie und perfekter Illusion und Täuschung ausmachen. Einmal zur Hand oder antrainiert erleichtern Techniken dabei ungemein das Metier der Scheinerzeugung. Effizienzgewinne durch verbesserte Techniken gibt es auch auf dem Gebiet der Täuschungsproduktion. Und neue technische Entwicklungen und Tools eröffnen oft ebenso neue, verblüffende Möglichkeiten für Immersion und Irreführung.

Der Beziehung von Technik und Täuschung und Illusion widmet sich die diesjährige Ausgabe des Jahrbuch Technikphilosophie. Das Begriffspaar Täuschung und Illusion baut dabei eine bewusst breite Spannweite an Phänomenen auf. Zwar basiert deren Grundgrammatik stets auf dem Spiel mit der Differenz von Sein und Schein; die obigen Beispiele zeigen aber bereits an, dass dies variantenreich erfolgen kann: Manipulation, Fake, Lüge, Simulation, Irrtum, Einbildung, Fiktion oder auch nur die Steigerung von Anmutung, Anschein und Unsicherheit – all dies stellt unterschiedliche Konfigurationen von Täuschung und Illusion dar. Zu der Vielfältigkeit dieser Formen kommt die Vielfältigkeit der Motive, aus denen Menschen Irrtümer und Illusionen erzeugen; sei es aus manifester Täuschungsabsicht, aus einer bloßen Lust am Spiel mit der Illusion oder – vielleicht paradox – gerade zu Aufklärungszwecken.

Auch wenn sich vielleicht die Geschichte der Technik als eine Geschichte der Scheinerzeugung erzählen ließe, gegenwärtig drängt sich ihr Zusammenhang besonders auf. Aktuelle technische Entwicklungen eröffnen neue Möglichkeiten der Erzeugung von Schein und Lüge, die infrage stellen, was einst als gesichert originär, authentisch und wahr angenommen werden konnte. Dabei sind derzeit auf frappierende Weise gleich zwei Grundmedien betroffen, durch die wir als Menschen indirekt, nicht durch unmittelbare eigene Wahrnehmung überprüfbare, Informationen gewinnen können: Sprache und Bilder. Sprache ist dies durch die rasante Entwicklung auf

dem Gebiet der Large Language Models, die – wie viele schmerzhaft oder mit Amüsement feststellen konnten – sich entweder unfreiwillig von den Fakten entfernen und ›halluzinieren‹ oder gar gezielt zur Generierung geschickt manipulativer Texte eingesetzt werden; nicht zuletzt begegnet uns hier auch die tragende Grundillusion, dass hier ein Mensch zu schreiben scheint, wo in Wahrheit nur eine Maschine operiert. Ebenso kennt das Medium der Bilder durch die ebenfalls KI-getriebene und rasant fortschreitende Entwicklung auf dem Gebiet der Deepfakes neue Täuschungsmöglichkeiten, die Erstaunen hervorrufen – hier scheint der Realisierung visueller Fantasien keine Grenze mehr gesetzt, bis auf die Mühe, einen passenden Prompt einzutippen.

Dies ist nur ein Gebiet der aktuellen rasanten Entwicklungen scheinerzeugender Techniken. Mit ihnen aber schreitet auch die Entwicklung der Entlarvungstechniken voran, die den neuen Täuschungsformen entgegenwirken sollen. Auch hier spielen KI-Anwendungen die entscheidende Rolle und ermöglichen neue Wege der direkten Lügendetektion, der Detektion von Fake News oder Deepfakes. Deren Entwicklung und Einsatz löst nicht einfach ein Problem, es wirft selbst wiederum neue auf – keine rein technischen –, sondern ebenso ethische, rechtliche und politische Probleme. Zunehmend ist wahrnehmbar, dass der Streit um Entlarvungstechniken ein politischer, zum Teil sehr polarisierter Streit ist.

Der ernsten Seite der technischen Scheinerzeugung, dort, wo sie uns ungewollt und schadhaft täuscht, steht die »leichte« Seite gegenüber – dort, wo Technik die *Hingabe* an die Fiktion fördert: Die stilistischen Techniken eines packenden Romans, die Schauspieltechniken auf der Bühne, oder graphische Techniken in Computerspielen. Auch Fantasien sind kein technikfreier Raum; sie können technisch angeregt und durch Techniken gelenkt werden.

Bei dieser schroffen Gegenüberstellung zwischen einerseits schadhaften Täuschungen, die einem andere antun und die wir unwissentlich und unwillentlich erleiden, und andererseits dem freiwilligen und lustvollen Sich-Einlassen auf Illusionen sollte ein dazwischenliegender Mischbereich nicht aus den Augen geraten. Besonders perfide ist die Täuschung dort, wo sie zur *Manipulation* wird, sprich dort, wo sie gerade ein Moment der Freiwilligkeit der manipulierten Person ausspielt, wo sie vermeintlich nicht übergriffig ist und sich bewusst harmlos gibt. Gerade diese Form aber scheint eine besondere technische Raffinesse und Geschicklichkeit zu erfordern, um nicht durchschaubar zu sein.

Man sieht schnell: Das Feld an Phänomenen im Bereich von Täuschung und Illusion ist breit. Und: Alle kennen Techniken zur Ermöglichung, Verbesserung und Vereinfachung derselben. Ein nur kleines Gebiet auf diesem Feld wollen wir im diesjährigen Jahrbuch abstecken.

Der erste Beitrag zum Schwerpunkt *Wer ist hier eigentlich Fake News?: Automatisierte Fake News Erkennung und die Politik des öffentlichen Raumes* von **Jörn Wiengarn** widmet sich einer Entlarvungstechnik – in diesem Fall: der automatisierten Fake News Detektion. Der Beitrag zeigt verschiedene Hinsichten auf, in denen es diffizil sein kann, Nachrichten als Fake News zu klassifizieren. Im Lichte dieser Einordnungsschwierigkeiten interpretiert der Autor die Klassifikationsfrage als in einem normativen Spannungsfeld situiert, das im Kern die Werte der Inklusivität sowie der epistemischen Qualität des öffentlichen Raumes betrifft. Diese normative Spannung ist nicht nur grundlegend für die Praxis der Content-Moderation überhaupt, sondern, so die zentrale These des Beitrags, schreibt sich auch invisibel in Detektionstechniken für Fake News ein.

Der Beitrag *Täuschung im Grenzgebiet: Erkenntnistheoretische Prüfung automatischer Betrugserkennung* von **Daniel Minkin** befasst sich kritisch mit dem Einsatz von automatisierten Technologien zur Lügendetektion. Im Fokus steht hierbei das ADDS – *Automatic Deception Detection System* –, das im Rahmen von iBorderCtrl entwickelt und innerhalb der EU eingesetzt wurde (und vielleicht noch wird); es handelt sich dabei um ein von der Europäischen Union gefördertes Forschungs- und Entwicklungsprojekt, das der effizienteren Kontrolle der EU-Außengrenzen dient. Das technische Modul ADDS stellt dabei eine Art KI-Lügendetektor dar, mittels dessen die Aussagen von einreisenden Personen daraufhin überprüft werden sollen, ob sie andere täuschen wollen. Welche Kritiken gegenüber dem ADDS wurden bislang vorgebracht? Worin liegen die Defizite der bisher vorgebrachten Kritik und an welcher Stelle ließen sie sich vertiefen? Zur Beantwortung dieser Fragen zieht Minkin Einsichten aus der erkenntnis- und wissenschaftstheoretischen Diskussion heran.

Geht es in den ersten beiden Artikeln um Technologien, mittels derer man zielsicherer die Täuschungsabsichten von Menschen aufdecken soll, so nimmt der Artikel *Die Heranzüchtung künstlicher (Ver)Sprecher: Dennett und Brandom über das Beherrschen inferentieller Fähigkeiten von sprachbasierter KI* von **Sebastian Bücker** speziell Technologien in den Blick, die den Schein einer sprechenden Person erzeugen, hier: ChatGPT. Ausgehend von der Beobachtung, dass sogenannte Large Language Models (LLMs) Texte produzieren können, die den Anschein geben, als seien von einer Person

verfasst, wirft der Artikel die grundlegende Frage auf, wie die sprachlichen Fähigkeiten von LLMs nun eigentlich zu verstehen sind. Versteht ChatGPT den von ›ihm‹ generierten Text? Oder produziert es hier – gewissermaßen blind – lediglich eine symbolische Struktur? Was heißt Verstehen? Daniel C. Dennetts und Robert B. Brandoms Überlegungen zum kompetenten Sprechen dienen hierbei dem Autor als sprachphilosophische Grundlage.

Im Beitrag *Die Metapher des Holografischen* nimmt **Jens Schröter** Holografie nicht als technische Bezeichnung, sondern als Metapher in den Blick und stellt einige ihrer Funktionen heraus. Er verortet diese u.a. im Rahmen eines medienhistorischen Narrativs, das auf die Idee einer »totalen Illusion« zuläuft. Losgelöst von ihrer technischen Bedeutung zeigt der Autor, wie das Hologramm in wissenschaftlichen wie populären Diskursen metaphorisch aufgerufen wird – häufig, um einen futuristischen Anstrich zu erzeugen oder Versprechen von Grenzüberschreitungen zu evozieren. Insgesamt erscheint Holografie als kulturelles Raster, das zur Stabilisierung kapitalistischer Fortschrittserwartungen beiträgt.

Um Immersion – genauer: das Verschwimmen von Realem und Inszenierten – geht es in **Theresa Schütz'** Beitrag *Penetrante Affizierung: Bindungstechnologien in immersiven Lebensformen am Beispiel der Muehl-Kommune (1974–1986)*. Ausgehend vom Konzept des immersiven Theaters nimmt Schütz die Praktiken der Muehl-Kommune in den Blick, die dazu beigetragen haben, die Kommunard:innen an die Regeln und Ideale der Kommune zu binden. Diese verpflichtete sich dem Ideal einer freien Liebe und verstand sich als Alternative zur Kleinfamilie. Sie geriet jedoch in heftige Kritik, als zahlreiche sexuelle Missbrauchsfälle an Minderjährigen innerhalb der Kommune bekannt wurden.

In der Rubrik *Abhandlungen* befasst sich **Dawid Kasprowicz** mit *Virtual Reality in der Wissenschaft*. Anders als im Schwerpunkt geht es ihm bei diesem Thema nicht um Täuschung und Illusionen, sondern vielmehr um die Frage, welchen methodischen und epistemischen Status VR hat, wenn sie als wissenschaftliches Mess- und Visualisierungswerkzeug genutzt wird.

Das diesjährige Archiv widmen wir dem spanisch-mexikanischen Technikphilosophen **José Gaos González Pola**. **Nicole C. Karafyllis** hat Ausschnitte aus Gaos' Text *Über die Technik* (*Sobre la técnica*) von 1959 ausgewählt und kommentiert sowie eine kurze kontextualisierende Einführung in Leben und technikphilosophisches Werk des Autors vorangestellt. Karafyllis verortet Gaos als gegenüber Ortega y Gasset und Heidegger eigenständigen Technikphilosophen, der sich unter anderem durch einen phänomenologischen Blick auszeichnet, der stets auch kritisch die lebens-

weltlichen Auswirkungen von Technik einfängt. Dem Archivbeitrag nachgestellt sind *Reflections on the 'Wirkungsgeschichte' of José Gaos' Philosophy of Technique*, in denen **María Antonia González Valerio** die Wirkungsgeschichte von Gaos einordnet. Valerio zeichnet die Trajektorien von Gaos' Schüler:innen nach, erläutert seine Philosophie vor dem Hintergrund der mexikanischen Exilsituation und stellt heraus, inwiefern sein Schaffen insgesamt von einem Hinwenden von der Transzendentalphänomenologie zu historisch-kontextualistischen Reflexionen geprägt ist.

In der Rubrik *Diskussion* rezensiert **Sven Thomas Mark Coeckelberghs** und **David J. Gunkels** Buch *Communicative AI: A Critical Introduction to Large Language Models*. Thomas hebt dabei einerseits den philosophisch klaren Charakter als kritisches Einführungsbuch hervor, das andererseits vor einer klaren politischen Positionierung zugunsten strengerer Regulierung von LLMs nicht zurückscheut. Kritisch sieht Thomas hingegen insbesondere Thesen, wie die eines Endes der Vorherrschaft gesprochener Sprache – sowohl empirisch als auch im Kontext des politischen Anspruchs des Buches. Ebenso kritisch führt er aus, dass das Verständnis von Autorschaft innerhalb der Mensch-Maschine-Interaktion mit LLMs im Buch doppeldeutig und unklar bleibt.

Kai Denker befasst sich in seiner Rezension mit **Rainer Mühlhoffs** Buch *Künstliche Intelligenz und der neue Faschismus*. Er geht hierin insbesondere der Frage nach, ob Mühlhoffs Faschismusbegriff geeignet ist, um die derzeit demokratiegefährdenden Tendenzen in Hinblick auf den Gebrauch von KI in den USA und Europa angemessen zu beschreiben. Vermag Mühlhoffs Faschismusbegriff Neues in den Blick zu nehmen, oder handelt es sich hierbei eher um einen politischen Kampfbegriff, der lediglich dazu dienen soll, die Dringlichkeit der Thematik zu unterstreichen? Denker zeigt auf, dass eher letzteres der Fall ist, wenngleich er das Anliegen des Buches teilt.

Bruno Gransche rezensiert die Monographie *Technikphilosophie: Neue Perspektiven für das 21. Jahrhundert* von **Kevin Liggieri**, **Marco Tamborini** und **Olivier Del Fabbro**. Gransche hebt positiv den informativen und anregenden Überblick über zentrale Problemfelder in der gegenwärtigen Technikphilosophie hervor, die anhand geschickt gewählter Begriffspaare verfährt und dabei insgesamt Technikphilosophie als kritische Übung praktiziert. Kritisch merkt Gransche hingegen insbesondere die Spannung zwischen dem Anspruch auf philosophischem Tiefgang und dem knappen Überblickscharakter an, ebenso fehlende Erläuterungen einiger Grundbegriffe sowie fehlende Bezüge zur internationalen Technikphilosophie.

Christian Driesen bespricht **Manfred Sommers** Buch *Stoff, Blatt und Kant. Philosophie des Graphismus.* Er hebt das Singuläre und Neuartige von Sommers phänomenologischen Ansatz hervor, unsere leiblichen Selbstverhältnisse von der graphischen Tätigkeit (Kritzeln, Zeichnen, Schreiben etc.) her zu erschließen. Nicht nur nimmt er hierbei die Bewegung der Hand als die linienzeichnende Instanz in den Blick, sondern auch das leere Blatt, das vermittelst dieser Leere ein Eigenleben führt und gewissermaßen die Hand zum Zeichnen bewegt.

Jessica Hainke rezensiert das Buch *Diskursive Unruhe. Max Bense in der Nachkriegsära (1945–1956)* von **Masetto Bonitz**, das aus vorrangig literaturwissenschaftlicher Perspektive das Zusammenwirken der biographischen Situation Benses mit seinem Werk beleuchtet. Hainke hebt insbesondere die von Bonitz erschlossenen technikphilosophischen Bezüge Benses hervor. Sie merkt kritisch an, dass ein Gesamtüberblick über den recht unbeforschten Nachlass Benses fehlt und dass eine Begründung der konkreten Quellenauswahl wünschenswert gewesen wäre.

Gegenstand der diesjährigen *Kontroverse* ist die Rolle der Technikphilosophie bei der philosophischen Durchdringung von KI. Die Bedeutung und Bewertung dieser Technologie ist nicht mehr bloß Thema spezialisierter technikphilosophischer und -historischer Beiträge; in dem Maße, in dem diese Technologie alle Lebensbereiche zu durchdringen und zu betreffen scheint, ist sie auch in allen philosophischen Binnendisziplinen (Ethik, Wissenschaftsphilosophie, Rechts- und politischer Philosophie) aufgegriffen worden. Dies ist Anlass, zwei Fragen zu stellen: In welcher Weise wird die *technische* Dimension von KI in den philosophischen Binnendisziplinen verhandelt? Und ist für diese auch eine spezifisch technik*philosophische* Sichtweise prägend oder relevant? Zu diesen Fragen positionieren sich **Johannes Lenhard, Florian J. Boge, Jan Georg Schneider, Anna Strasser, Karoline Reinhardt** und **Susanne Beck**.

In seinem *Kommentar The Cloud is on the Ground: the Magic Tricks of Silicon Valley* geht **Gerry McGovern** den Strategien des Silicon Valley nach, die materielle Dimension von Digitalität aufgrund ihrer fraglichen sozialen und ökologischen Implikationen zum Verschwinden zu bringen. Hinter einer schönen Fassade gerät schnell aus dem Blick: Nicht nur betreibt das Valley in hohem Maße ›Greenwashing‹, auch verlagert es umweltbelastende Technikproduktion in ärmere Regionen der Welt.

Zum Schluss widmet sich die Glosse von **Lucy Osler** und **Joel Kruger** dem Phänomen des AI Gossip. Dass KI-Chatbots zuweilen halluzinieren und Bullshit verbreiten können, ist hinlänglich klar. Ebenso klar ist, dass

KI falsche Tatsachen verbreiten kann, die öffentliche Personen zu Unrecht in ein schlechtes Licht rücken. Aber gibt es da nicht noch mehr: Kann KI nicht sogar geradezu lästern? Teilt sie mit Usern gerne pikante Tatsachen und Meinungen über andere? Und wenn ja: Was macht ›Gossip‹, wenn es von einem scheinbar nicht in unseren sozialen Beziehungen verwickelten Chatbot kommt?

Schwerpunkt

Jörn Wiengarn

»Wer ist hier eigentlich Fake News?«

Automatisierte Fake News-Erkennung und die Politik des öffentlichen Raumes

Abstract

Techniken der automatisierten Fake News-Erkennung (FND) sind ein verbreitet eingesetztes Hilfsmittel der Content-Moderation. In dem Artikel geht es darum aufzuzeigen, dass diesen FND-Techniken irreduzibel eine bestimmte Politik des öffentlichen Raumes eingeschrieben ist. Dafür geht der Artikel in drei Schritten vor: Zunächst fächert er verschiedene Hinsichten auf, in denen die Frage, welche Nachrichtenmeldungen als Fake News einzuordnen sind (und welche nicht) epistemisch unklar sein kann. Hier werden verschiedene Typen von Grenzfällen von Fake News aufgezeigt. Darauf aufbauend wird argumentiert, dass sich die Einordnung als ›fake‹ oder ›nicht-fake‹ auch als Priorisierung eines bestimmten nicht-epistemischen Wertes des öffentlichen Raums verstehen lässt. Genau genommen, soll nachgewiesen werden, wie sich diese Einordnungsfrage als situiert in der Spannung zwischen dem normativen Anspruch der Inklusivität des öffentlichen Raums vs. der Vernunft- und Wahrheitsorientierung desselben verstehen lässt. Nach einer Erläuterung der typischen Funktionsweisen der algorithmusbasierten FND-Techniken wird zuletzt aufgezeigt, inwiefern sich diese ebenso irreduzibel in dieser Spannung positionieren. Dies hängt damit zusammen, dass mit den Daten anhand derer die eingesetzten Algorithmen trainiert werden, bereits die wertgeladene Einordnungsfrage nach ›fake‹ oder ›nicht-fake‹ beantwortet sein muss.

Automated Fake News Detection (FND) techniques are a commonly used tool in content moderation. This article aims to demonstrate that these FND techniques are inextricably linked to a specific politics of the public sphere. The article proceeds in three steps: Firstly, it explores various ways in which the question of which news reports should be classified as fake news (and which should not) can be epistemically unclear. Here, different types of borderline cases of fake news are discussed. Building on this, it is argued that classifying a report as ›fake‹ or ›not fake‹ can also be understood as prioritizing a certain non-epistemic value of the public sphere. More precisely, this classification issue is shown to be situated within the tension between the normative requirement of the inclusivity of the public sphere vs. its orientation towards reason and truth. After explaining the typical functionalities of algorithm-based FND techniques, the article finally demonstrates how these techniques are also irreducibly positioned within this tension. This is related to the fact that the data used to train the algorithms already must answer the value-laden classification question of ›fake‹ or ›not fake‹.

Das Täuschungspotenzial von Fake News hat einen gleich doppelten Problemcharakter. Zum einen können Fake News durch das Imitieren von eigentlich glaubwürdigen Nachrichtenformaten Leser:innen zu falschen Überzeugungen führen. Zum anderen können sie aber auch Verwirrung stiften. Dadurch, dass sie glaubwürdige Nachrichten imitieren, *exploitieren*

sie schließlich auch deren Kredibilität. Einfach gesagt, provozieren Fake News dadurch auch eine höherstufige Frage, nämlich: Was kann man *überhaupt noch* glauben – und was nicht?

Informative Verwirrung und Orientierungslosigkeit sind somit die weiteren Folgen, die durch Fake News drohen, ebenso wie die durch sie gefütterte Segmentierung und Polarisierung öffentlicher Debatten.[1] Naheliegenderweise evoziert das Phänomen ›Fake News‹ deshalb auch ganz grundlegende demokratietheoretische Bedenken: Es scheint, sie drohen rationalen öffentlichen Diskursen immer stärker eine vernünftige Bezugsbasis zu entziehen. Nicht nur dort, wo Fake News und Desinformation in gezielten Kampagnen auf die Delegitimierung demokratischer Institutionen abzielen, zeigt sich deshalb ihre demokratieschädliche, wenn nicht gar -gefährdende Tendenz.[2] Kurzum, Fake News drohen eine sozialepistemologisch tiefgehende und gesamtgesellschaftlich bedeutsame Strukturkrise des Informations- und Diskursraumes zu befeuern.[3] Sie sind nicht nur ein Problem, da sie Individuen täuschen, sondern auch, da sie die *Öffentlichkeit als Ganze* dysfunktional zu machen drohen.

Die schwerwiegenden Folgen von Fake News drängen die Frage auf, wie es zu ihrer aktuellen Konjunktur kommen konnte. Hier nun scheint es, kann man über technische Entwicklungen als Teil der Erklärung nicht schweigen. Immerhin gibt es Fake News zwar im gewissen Sinne schon, solange es das Medium Zeitung gibt (das Phänomen als solches ist also historisch mitnichten neu).[4] Besonders akut scheint das Problem aber heute aufgrund technischer Neuerungen, die sowohl das Verfassen als auch

1 Zu einer Konzeptualisierung dieser Mechanismen als verschiedene Umstrukturierungseffekte von Vertrauensnetzwerken, siehe: Jörn Wiengarn und Maike Arnold: »Algorithm-Driven Reconfigurations of Trust Regimes. An Analysis of the Potentiality of Fake News«, in: Juliane Jarke u.a. (Hg.): *Algorithmic Regimes. Methods, Interactions and Politics*, Amsterdam 2024, S. 169–185.

2 Siehe exemplarisch für eine solche Problematisierung Lejla Turčino und Mladen Obrenović: *Fehlinformationen, Desinformationen, Malinformationen. Ursachen, Entwicklungen und ihr Einfluss auf die Demokratie*. Publikation der Heinrich Böll Stiftung, E-paper. 2020. Siehe auch prominent Jürgen Habermas: *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*, Berlin 2022.

3 Vgl. Axel Gelfert: »Fake News, False Beliefs, and the Fallible Art of Knowledge Maintenance«, in: Amy K. Flowerree u.a. (Hg.): *The Epistemology of Fake News*, Oxford 2021, S. 310–333.

4 Vgl. Lee McIntyre: *Post-Truth*, Cambridge/London 2018, S.70 – 73. Ebenso: Catarina Dutilh Novaes und Jeroen de Ridder: »Is Fake News Old News?«, in: Amy K. Flowerree u.a. (Hg.): *The Epistemology of Fake News*, Oxford 2021, S. 156–179.

das Verbreiten von Fake News massiv erleichtern. So ist mit dem Internet nicht nur ein digitaler Informationsraum entstanden, der insgesamt die Menge an verfügbaren Informationen hat explodieren lassen. Insbesondere schafft er Räume, in denen die alte massenmediale Logik der Trennung zwischen Autor:innen und Rezipient:innen von öffentlichen Informationen aufgehoben wurde.[5] Noch nie ging es so schnell und einfach, eigene Inhalte zu kreieren und auch in Form von Fake News an eine große Öffentlichkeit zu bringen – vorbei an den Institutionen des klassischen Journalismus, denen rückblickend die Funktion von ›Gatekeepern‹ zugeschrieben werden kann.[6] Hinzu kommen neue KI-Systeme wie GPT, die das Generieren von Fake News beschleunigen, sowie insgesamt eine technisch-algorithmische Infrastruktur der sozialen Medien, die deren Viralität begünstigt – und damit Fake News ein großes Stück der Aufmerksamkeitsressourcen zukommen lässt.[7]

Neue technische Möglichkeiten erhöhen die Masse an Fake News, zugleich liegt in der Technik aber auch ein Versprechen, dieser Masse Herr zu werden. Schließlich droht die schiere Flut an Fake News in den sozialen Medien menschliche Faktenchecker schlicht zu überfordern, oder lässt diese nicht mehr so schnell hinterherkommen, wie es eigentlich vonnöten wäre, um die Viralität von Fake News möglichst früh einzudämmen. Nicht zuletzt versprechen automatisierte Methoden der Fake-News-Detektion (im Folgenden: FND) damit auch erhebliche Kosten einzusparen.[8] Diese werden deshalb als entscheidendes Komplement in der Content-Moderation, und damit als unverzichtbarer Teil der Lösung des Fake News-Problems angesehen.[9] Möglichkeiten der Automatisierung von Content-Moderation erscheinen deshalb besonders attraktiv. Nicht überraschend boomt deshalb auch das Feld der automatisierten Fake-News-Detektion in der Informati-

5 Vgl. Habermas: *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*, S. 44–47. Sue Robinson und Cathy DeShano: »›Anyone Can Know‹. Citizen Journalism and The Interpretive Community of The Mainstream Press«, in: *Journalism* 12 (2011), Heft 8: S. 963–982.

6 McIntyre: *Post-Truth*, Kap. 4 und 5.

7 Für eine Übersicht darüber wie aktuelle technische Entwicklungen und Implementationsentscheidungen die Verbreitung von Fake News erhöhen, siehe: Noah Giansiracusa: *How Algorithms Create and Prevent Fake News*, New York City 2021.

8 Vgl. Tarleton Gillespie: »Content Moderation, AI, and the Question of Scale«, in: *Big Data & Society* 7 (2020), Heft 2, S. 1–5.

9 Vgl. Victoria L. Rubin: *Misinformation and Disinformation. Detecting Fakes with the Eye and AI*., Cham 2022, Kap. 8.1.1.

onstechnik, im Besonderen im ML-Bereich.[10] Dabei werden Algorithmen eingesetzt, die versprechen Fake News großflächig schneller zu erkennen. Je nach Implementationsweise werden derart identifizierte Fake News entweder automatisiert bearbeitet, oder aber, wie es meistens geschieht, an menschliche Faktenchecker:innen zur endgültigen Überprüfung weitergeleitet, und dann je nach Verfahrenspraxis entweder mit einem Warnhinweis oder zusätzlichen Informationen versehen, in der Sichtbarkeit reduziert oder schlichtweg gelöscht.[11]

So verführerisch dieses von der Technik ausgehende Versprechen klingen mag, so sehr täuscht es aber über diffizile vorgelagerte politische Fragen hinweg. Wie schließlich die zentrale These lautet, die in diesem Artikel plausibilisiert werden soll, ist den Detektionsalgorithmen unvermeidlich eine bestimmte Politik der Öffentlichkeit eingeschrieben. Grundsätzlich gilt, dass die Frage danach, was letztlich als Fake News anzusehen ist, nicht trivial ist. Nichttrivial ist diese Frage nicht nur deshalb, sofern sie epistemisch herausfordernd sein kann, sondern da es klassifikatorische Demarkationsprobleme gibt, die zutage treten lassen, dass die Frage, was als ›fake‹ anerkannt wird, *auch eine Wertenscheidung* ist – eine Entscheidung, die sich in Bezug auf in Spannung stehende Grundwerte des öffentlichen Raumes verhalten muss. Eine solche Wertpositionierung, so die hier zu entwickelnde These, schreibt sich implizit in die FND-Technik ein. Sie ist somit inhärent (und nicht bloß mit der Frage ihrer Anwendung) auf die eine oder andere Art wertgeladen. Wie ich zeigen möchte, *verkörpern FND-Techniken somit eine gewissermaßen invisible Politik des öffentlichen Raumes*.

Im Einzelnen werde ich wie folgt vorgehen. In einem ersten Schritt werde ich die prinzipiellen Klassifikationsschwierigkeiten von Fake News erläutern und hier systematisch verschiedene Ebenen der klassifikatorischen Graubereiche unterscheiden. Zweitens werde ich argumentieren, dass die Entscheidung, was als ›fake‹ oder ›nicht fake‹ eingeordnet wird, aus der praktischen Perspektive der Content-Moderation als wertgeladene Entscheidung aufzufassen ist. Im Kern besteht hier eine Spannung zwischen zwei normativen Anforderungen an eine demokratische Öffentlichkeit, nämlich deren Inklusivität, im Sinne der Wahrung des Partizipationsrech-

10 Alessandro Bondielli und Francesco Marcelloni: »A Survey on Fake News and Rumour Detection Techniques«, in: *Information Sciences* 497 (2019), S. 38–55, hier S. 39.

11 Zur Übersicht aber auch kritischen Auseinandersetzung mit diesem Versprechen der Technik, siehe: Gillespie: »Content Moderation, AI, and the Question of Scale«.

tes an Meinungs- und Willensbildungsprozessen, einerseits und deren Vernunft- und Wahrheitsorientierung andererseits. Im Raum dieses Konflikts sind auch FND-Algorithmen zu verorten. Um dies zu zeigen, werde ich, drittens, einen Überblick über die prinzipielle Funktionsweise der hierbei eingesetzten Algorithmen geben. Daran anschließend werde ich zuletzt konkret analysieren, inwiefern technische Entscheidungen hier implizite Wertentscheidungen im zuvor erläuterten Sinne widerspiegeln. Dies wird primär mit Verweis auf die Verwendung der Daten geschehen, die zum Trainieren, Optimieren und Evaluieren der Algorithmen herangezogen werden.

»I know it when I see it«? Die klassifikatorischen Graubereiche von Fake News

Was sind eigentlich Fake News? Wenn man diese Frage begriffsklärend versteht, mangelt es in der philosophischen Diskussion nicht an Antwortvorschlägen: Fake News können verstanden werden als bloße Falschmeldung, als Lügen mit Täuschungsabsicht der Verfasser oder als spezifische Form von ›Bullshit‹.[12] Es gibt zwar Versuche, eine allgemeingültige und klar hergeleitete Definition von Fake News zu formulieren, allerdings kann man nicht davon sprechen, dass sich eine solche durchgesetzt hätte.[13] Insbesondere umstritten ist dabei die Frage nach den Motiven der Produzent:innen von Fake News: Gehört zu Fake News qua Definition auch, dass ihre Ver-

12 Für eine Übersicht über die Definitionsangebote verweise ich hier nur auf: Romy Jaster und David Lanius: »Speaking of Fake News. Definitions and Dimension«, in: Amy K. Flowerree u.a. (Hg.): *The Epistemology of Fake News*, Oxford 2021, S. 19–45. Ebenso liefern Tandoc, Lim und Ling auf Grundlage einer multidisziplinären Literaturauswertung eine typologische Übersicht über typische Fake News-Definitionen. Siehe Edson Tandoc, Zheng Wei Lim und Richard Ling (2018) »Defining ›Fake News‹. A Typology of Scholarly Definitions«, in: *Digital Journalism 6* (2018), Heft 2, S. 137–153. Sowohl Jaster und Lanius als auch Tandoc et al. streichen dabei zwei ähnliche zentrale Dimensionen hervor, entlang derer sich die Definitionen unterscheiden können: Zum einen anhand des Faktizitätslevels, damit ist grob das Maß gemeint, in dem sich Fake News auf wahre Sachverhalte beziehen. Zum anderen anhand der Motive der Verfasser von Fake News.

13 Siehe auch zur grundlegenderen methodischen Kritik an den Versuchen zur Bestimmung einer kontextübergreifend ›richtigen‹ Definition von Fake News durch Explikation durch Sprachintuitionen: Thomas Grundmann: »Fake News: The Case for a Purely Consumer-Oriented Explication«, in: Inquiry 66 (2020), Heft 10, S. 1758–1772. Ebenso: Frank Hofman: »Fake News und unsere epistemische Grundsituation«, in: Amelie Bendheim und Jennifer Pavlik (Hg.): *›Fake News‹ in Literatur und Medien.*

fasser:innen kein Interesse an deren Wahrheitsgehalt haben? Und wenn ja, zeichnen sich Fake News dadurch aus, dass hinter ihnen eine Täuschungsabsicht steckt? Oder sind Fake News durch ein bloßes Desinteresse an der Wahrheit seitens ihrer Verfasser:innen charakterisiert?

Aufgrund der Schwierigkeit, dass auf keine allgemeinverbindliche Definition von Fake News verwiesen werden kann, scheint es umso sinnvoller mit einem für unsere Zwecke zugeschnittenen Minimalbegriff von Fake News zu arbeiten, der den Kern des Problemcharakters von Fake News für den öffentlichen Raum hervorhebt – der Problemcharakter der Fake News auch zu einem potenziellen Ziel der Content-Moderation macht. Er besteht im Grunde darin, dass man es mit öffentlichen Nachrichteninhalten zu tun hat, die einerseits dadurch glaubwürdig erscheinen, dass sie nachrichtenförmig aufbereitet sind (1), in Wirklichkeit jedoch falsche oder irreführende Inhalte verbreiten (2). Dieser unumstrittene Problemkern soll im Folgenden auch die Definition von Fake News darstellen, von der ich ausgehen werde.[14]

Zur kurzen Erläuterung:

1. Fake News versuchen sich als glaubwürdige Nachrichtenformate auszugeben. Illokutionär betrachtet lassen sie sich damit ganz grundlegend als Kommunikationselemente mit einer Wahrheitsgarantie einordnen.[15] Allein durch ihr Erscheinen im Format einer *Nachricht* haftet ihnen ein

---

Fakten und Fiktionen im interdisziplinären Diskurs. Bielefeld 2022, S. 35–54, hier S. 35f.

14 Die Definition entspricht damit dem, was Zhou u.a. für die informationstechnische Literatur als weiteste Definition von Fake News ansehen, siehe: Xinyi Zhou und Reza Zafarani: »A Survey of Fake News. Fundamental Theories, Detection Methods, and Opportunities«, in: *ACM Computing Surveys* 53 (2020), Heft 5, S. 1–40, hier S. 4. Mit guten Gründen nehmen viele Definitionen (auch in der technischen Literatur) an, dass mit Fake News eine Täuschungsabsicht oder zumindest ein Desinteresse ihrer Verfasser:innen an der Wahrheit verbunden sei. Ich blende diesen Aspekt hier aber aus, da er zumindest nicht die Eigentlichkeit des Problems von Fake News auszumachen scheint. Unabhängig von den Verfasser:innen nämlich scheinen täuschende oder irreführende Nachrichten *als solche* zunächst ein Problemartefakt für die Öffentlichkeit zu sein. Hinzu kommt, dass die Absicht als solche auch für die Techniken der automatisierten Fake News Erkennung zumindest keine direkte Rolle spielen, da sich diese schwer direkt identifizieren lässt.

15 Vgl. Jaster: »Speaking of Fake News«, in: Floweree (Hg.): The Epistemology of Fake News, S. 20; Nicola Mößner: »Trusting the Media? TV News as a Source of Knowledge«, in: *International Journal of Philosophical Studies* 26 (2018), Heft 2, S. 205–220, hier: S. 212. Ebenso wird dies umgekehrt auch auf der Rezipientenseite von Nachrichten erwartet, siehe: Bill Kovach und Tom Rosenstiel*: The Elements of Journalism:*

assertorischer Charakter an, denn Nachrichten erheben der Textart nach den Anspruch darauf, etwas Neues, öffentlich Relevantes *und Wahres* mitzuteilen. Die Nachrichtenförmigkeit hat dabei auch die Funktion, an ein grundlegendes Vertrauen der Rezipient:innen in Standard-Nachrichtenformate anzuknüpfen und dieses zu exploitieren. Erst dadurch sind sie überhaupt in der Lage ihren in der Einleitung beschriebenen doppelten Problemcharakter anzunehmen: Durch das Nachahmen vermeintlich seriöser Nachrichtenformate erhalten sie zum einen eine scheinbare Glaubwürdigkeit, die ihren *täuschenden* Charakter ausmacht. Zum anderen können Fake News dadurch auch das Vertrauen, das tatsächlich wahrhaftigen Nachrichtenmeldungen entgegengebracht wird, langfristig unterminieren oder zumindest irritieren.[16]

2. Im Gegensatz zu glaubwürdigen Nachrichten sind Fake News falsch oder irreführend. Der Begriff ›irreführend‹ umfasst dabei variantenreiche Formen, in denen Fake News zwar nicht wortwörtlich etwas Falsches sagen, aber etwas Falsches suggerieren oder implizieren. Jaster und Lanius beispielsweise weisen auf den Fall hin, in dem Breitbart im Januar 2017 darüber berichtete, wie eine Gruppe von etwa 1.000 Menschen in Dortmund eine historische Kirche in Brand steckten.[17] Wortwörtlich gesehen stimmt diese Aussage, da in der Silvesternacht aus Versehen ein Sicherheitsnetz am Gerüst einer Kirche durch Feuerwerkskörper entzündet wurde. Die Meldung suggeriert aber mindestens darüber hinaus, dass dies mit kollektiver Absicht geschehen sei und kann insofern als Fake News angesehen werden.

Damit stellen Fake News insgesamt eine spezifische Unterkategorie von Falschinformation allgemein dar. Ihr abgrenzendes Spezifikum liegt dabei in ebenjenem erstgenannten Aspekt: Fake News sind eine Form von Falschinformation, die per definitionem in ihrem Erscheinungsbild seriöse Nachrichtenformate nachahmen.

What Newspeople Should Know and The Public Should Expect. New York 2007, hier: S.17.

16 Zu diesem doppelten Problemcharakter siehe Sanford C. Goldberg: »Fake News and Epistemic Rot; or, Why We Are All in This Together«, in: Amy K. Flowerree u.a. (Hg.): *The Epistemology of Fake News*, Oxford 2021, S. 265 -285.

17 Jaster: »Speaking of Fake News«, in: Floweree (Hg.): *The Epistemology of Fake News*, S. 21. Wie Jaster und Lanius hervorheben wird hier ausgelassen, dass dies unabsichtlich geschehen sei. Tatsächlich haben die Feuerwerkskörper aus Versehen ein Fangnetz an einem Baugerüst an der Kirche zum Brennen gebracht. Das Feuer war nach Feuerwehrangaben klein und konnte unmittelbar gelöscht werden.

Man mag vielleicht an dieser Stelle einwenden, dass der Terminus Fake News aktuell zu stark politisch aufgeladen sei bzw. als politischer Kampfbegriff zur Delegitimierung einer politischen Gegenseite, und deswegen nicht sinnvoll als wissenschaftlicher Begriff zu verwenden sei.[18] Auch wenn ich die Kritik an der möglichen Verwendungsweise des Begriffs nachvollziehe, bleibe ich jedoch aus drei Gründen weiterhin bei dieser Vokabel: Zum einen, da ich den Begriff hier explizit unabhängig von seinen Verwendungsweisen als klar definierten *terminus technicus* einführe – der zudem in diesem Kontext als allgemeiner Genus und nicht als politischer Kampfbegriff zu verstehen ist.[19] Hinzu kommt aber auch, dass der Terminus den zentralen erstgenannten Aspekt zum Ausdruck bringt, nämlich, dass es sich hierbei speziell um vermeintliche Nachrichten handelt. Und drittens versuche ich schlichtweg eine pragmatische Kontinuität zu der gängigen Forschungsliteratur beizubehalten. Der Begriff Fake News hat sich, wenn auch in den klaren definitorischen Konturen flexibel, in Teilen der Forschungslandschaft als Fachbegriff eingebürgert. Er ist zudem in der genannten Forschungsrichtung der Fake-News-Detektion (FND) zentral, auf die ich mich unten beziehen werde.

Soweit zur Begriffsbestimmung von Fake News. Die Frage danach, was Fake News sind, lässt sich nun aber nicht nur definitorisch, sondern auch als klassifikatorische Herausforderung verstehen: Was fällt eigentlich *konkret unter* diesen Begriff – und was nicht. Dies jedoch lässt sich keinesfalls immer so eindeutig an den jeweiligen Nachrichtenmeldungen ablesen. Wie ich im Folgenden aufzeigen möchte, erweist es sich gleich auf mehreren Ebenen als nicht trivial, was als Fake News einzustufen ist, da sich hier Graubereiche auftun. Dies betrifft, so möchte ich hier vorschlagen, erstens eine illokutionäre, zweitens eine inhaltliche und drittens eine epistemische Ebene, sowie viertens das, was ich die Ebene der Gesamteinordnung nennen möchte:

18 Joshua Habgood-Coote: »Stop Talking about Fake News!«, in: Inquiry 62 (2019), Heft 9–10, S. 1033–1065; David Coady: »The Trouble with ›Fake News‹«, in: *Social Epistemology Review and Reply Collective* 8 (2019), Heft 10, S. 40–52.

19 Vgl. diese methodische Herangehensweise in: João Pedro Baptista und Anabela Gradim: »A Working Definition of Fake News«, in: *Encyclopedia* 2 (2022), S. 632–645, hier: S. 635. Für eine grundlegende begriffsstrategische Verteidigung siehe auch: Sven Bernecker, Amy K. Flowerree, und Thomas Grundmann: »Einleitung«, in: Amy K. Flowerree u.a. (Hg.): *The Epistemology of Fake News*, Oxford 2021, S. 1–16, hier S. 5f.

1. In der obigen Definitionserläuterung wurde festgehalten, dass Fake News illokutionär betrachtet assertorischen Charakter haben: Sprechakttheoretisch geben sie also eine Art Garantie auf die Wahrheit dessen, was sie behaupten. Tatsächlich aber gibt es eine ganze Reihe an Präsentationsmodi von Falschaussagen, die deren illokutionären Charakter verwischen. Hier lässt sich schlichtweg nicht eindeutig sagen, *wie* solche Aussagen gemeint sind – und ob mit ihnen eine *Garantie auf Wahrheit* verbunden ist. Oft ist dies sogar gerade Teil einer Unverbindlichkeitsstrategie, die in Desinformationskampagnen genutzt wird.
   Ein gutes Beispiel hierfür sind Aussagen, die sich als ›Meinungsäußerung‹ ausgeben. Bei diesen ist tendenziell unklar, ob sie mit einer Garantie auf Wahrheit einhergehen, oder sich nicht vielmehr expressiv, eben als bloßer *Ausdruck* einer Meinung verstehen lassen müssen. Aussagen können so formuliert werden, dass sie auf der feinen Linie zwischen Assertion und Expression balancieren. Hieraus erklärt sich auch die notorische Kontroverse darum, ob Meinungsbeiträge, auch wenn sie falsche Aussagen beinhalten, als legitime Beiträge zu bewerten sind, oder vielmehr problematische Schlupflöcher für eigentliche Desinformation darstellen.[20]
   Ein weiteres Beispiel für den illokutionären Graubereich sind Aussagen, die aufgrund ihres womöglich (halb-)ironischen oder sarkastischen Tonfalls schwer einzuordnen sein können. Üblicherweise (aber auch das ist nicht ganz unumstritten) wird Satire zwar nicht als Fake News eingestuft, zumindest insofern diese nicht zur ernsthaften Irreführung geeignet scheint. Aber auch hier gibt es einen Graubereich von Meldungen, bei denen – teilweise bewusst – vernebelt wird, ob diese tatsächlich ernst gemeint sind oder eine ernstgemeinte Aussage mit ernstgemeinter Wahrheitsgarantie bloß *nachahmen*.
2. Neben der Frage, *wie* eine Nachrichtenmeldung gemeint sein kann, kann aber auch uneindeutig sein, *was* genau die Meldung eigentlich besagt. Dies kann mitunter sehr direkt durch das Verwenden von mehrdeutigen Symbolen oder Ausdrücken geschehen.[21] Es kann aber auch elaboriertere

20 Diese Diskussion entzündete sich etwa an der Entscheidung Facebooks, Meinungsbeiträge vom Fact-Checking auszunehmen. Siehe: Jeff Horwitz: »Facebook to Exempt Opinion and Satire from Fact-checking«. *The Wall Street Journal*, 30.09.2019, https://www.wsj.com/articles/facebook-to-create-fact-checking-exemptions-for-opinion-and-satire-11569875314 (aufgerufen: 12.2.2024).

21 Anregend sind hierzu auch die Überlegungen von Jason Stanley, der verschiedene Bedeutungsebenen von Begriffen und Aussagen erläutert und wie diese in Propaganda

Formen annehmen. Die obige Definition von Fake News wies bereits darauf hin, dass ein großer Teil dessen, was in Nachrichten kommuniziert wird, nicht wortwörtlich formuliert, sondern vielmehr impliziert wird. In manchen Fällen mag aber nicht klar sein, ob eine Implikation vom Text offenbar ›mitgemeint‹ ist, oder vielmehr von der Leserin unzulässigerweise in diesen ›hineingelesen‹ wird. Auch hier gibt es offenbar keine klare Grenze und somit Fälle, die auf der Kippe stehen.

3. Auch wenn in einer Nachrichtenmeldung klar sein mag, wie und was diese meint, kann es immer noch Fälle geben, in denen aus Gründen der epistemischen Unsicherheit uneindeutig ist, ob diese als ›falsch‹ einzuordnen ist. So scheint es etwa berechtigt anzunehmen, dass die Leugnung des anthropogenen Klimawandels einen klaren Fall von Desinformation darstellt; und das, obwohl man zugestehen kann, dass es natürlich einen minimalen, prinzipiellen Restzweifel daran geben mag. Wo aber ist hier die Grenze zu ziehen? Ab welchem *Grad* an Gewissheit kann man hier eine Aussage eindeutig als falsch einsortieren? Auch in dieser Hinsicht gibt es offenkundig uneindeutige Grenzfälle.
4. Zuletzt gilt: auch in Fällen, in denen klar sein mag, was und wie etwas in einer Nachrichtenmeldung ausgesagt wird, und zudem relativ feststeht, was demgegenüber die tatsächliche Faktenlage ist, mag es weiterhin unklar sein, ob eine solche Meldung in der *Gesamtbetrachtung* als falsch eingestuft werden sollte. Wie etwa sind Fälle einzuordnen, bei denen in einer gesamten Nachrichtenmeldung nur wenige Falschaussagen vorkommen; oder wenn Sachverhalte schlichtweg nicht ganz ›sauber‹ dargestellt wurden; oder wenn Falschaussagen in Bezug auf belanglose Nebensächlichkeiten getroffen werden, die damit auch keinen großen Schaden zu verursachen drohen? Auch auf dieser Ebene tut sich ein Bereich der Unschärfe der Trennung zwischen ›fake‹ und ›nicht-fake‹ auf.

Die verschiedenen Arten von Grenzfällen zeigen somit auf, dass man keineswegs so einfach an den Meldungen *ablesen* könnte, ob diese letztlich als Fake News einzustufen sind.[22] Man kann hier gleichsam nicht nach dem

eingesetzt werden: Jason Stanley: *How Propaganda Works,* Princeton/Oxford 2015, Kap. 4.

22 Natürlich kann man argumentieren, dass dies doch gar nicht notwendig ist. Immerhin kann man auch nuanciertere Einordnungen vornehmen, eine Nachricht also als mehr oder weniger falsch bzw. irreführend einordnen. Tatsächlich wird dies üblicherweise auch im Bereich der Content-Moderation gemacht (s.u.). Das Grundproblem

Motto operieren: »I know it when I see it.«[23] Die Frage danach, wann Fake News eine illegitime Form der Täuschung darstellen, erweist sich in vielen Fällen vielmehr als sensibel und nicht durch klare Anwendung harter epistemischer Kriterien entscheidbar.

Allerdings, so werde ich im nächsten Schritt zeigen, muss die Frage der Einordnung auch gar nicht als ein rein epistemisches Klassifikationsproblem gedeutet werden. Im Lichte der Relevanz, die diese Unterscheidung schließlich für die Content-Moderation hat, zeigt sich vielmehr, dass solcherlei Klassifikationsfragen immer auch Fragen der Wertpriorisierung sind.

Fake oder Nicht-Fake? Die Politik einer Unterscheidung

Content-Moderation hat das allgemeine Ziel, Medieninhalte in Hinblick auf ihre Angemessenheit für eine entsprechende Plattform oder Webseite zu prüfen und im Falle der Nichtangemessenheit so zu intervenieren, dass deren Schaden begrenzt wird (im einfachsten Fall etwa, indem entsprechende Inhalte gelöscht werden). Medieninhalte können dabei aus verschiedenen Gründen unangemessen sein, etwa da es sich um Hassrede, Gewaltdarstellungen, politisch extremistische Inhalte oder illegale Inhalte handelt. Content-Moderation kann aber auch das Ziel haben Falschinformationen einzudämmen.

Aus Sicht der Content-Moderation stellt sich damit die Frage nach ›fake‹ oder ›nicht-fake‹ mit Blick auf einen ganz bestimmten Zweck, nämlich daraufhin, dass *problematische* Inhalte weniger oder gar keine öffentliche Beachtung erhalten sollen. Die Einordnung als ›fake‹ fällt also der Idee nach mit der Einordnung als für den öffentlichen Raum problematisch zusammen – mit der angestrebten Konsequenz, dass solcherlei Inhalte gelöscht, gekennzeichnet oder weniger sichtbar gemacht werden. Damit deutet sich an, dass die Frage der Einordnung als ›fake‹ oder ›nicht-fake‹ nicht einfach nur eine epistemische Frage darstellt, sondern immer auch

---

aber scheint zu bleiben, nämlich dass auch bei feineren Abstufungen Fälle auf der Kippe stehen können und die Frage der Klassifizierung uneindeutig bleibt.

23 Damit spiele ich an auf eine berühmte Aussage des Richters des Obersten Gerichtshofes der USA Potter Stewart zur Frage, wie man »hard-core pornography« erkennen könne. Siehe hierzu: Peter Lattman: »The Origins of Justice Stewart's ›I Know It When I See It‹«, in: *The Wall Street Journal*, 27.09.2007, https://www.wsj.com/articles/BL-LB-4558 (aufgerufen: 12.2.2024).

von einer normativen Vorstellung von der Gestaltung des öffentlichen Raumes geleitet ist bzw. sich hier auch normativ positioniert.

Einer ähnlichen Stoßrichtung folgend hat Elizabeth Stewart vorgeschlagen, die Entscheidung für oder gegen eine Klassifikation als ›Fake News‹ auch als Ausdruck der Priorisierung eines bestimmten Wertes des öffentlichen Raums zu verstehen. Genau genommen verortet sie diese Entscheidung in einen Konflikt zwischen zwei Werten, nämlich der freien Meinungsäußerung auf der einen Seite und dem Schutz der Öffentlichkeit vor unfairer und schädlicher Fehlinformation auf der anderen.[24] Dieser Idee folgend kann etwa Facebooks Entscheidung, Meinungsbeiträge prinzipiell vom Faktencheck auszunehmen, als Ausdruck der Priorisierung freier Meinungsäußerung verstanden werden;[25] die Entscheidung von Faktencheck-Webseiten wie Snopes hingegen, diese miteinzubeziehen, stellt sich wiederum als Ausdruck der Priorisierung von fairer und unschädlicher Kommunikation dar.

Der von Stewart beschriebene Grundkonflikt lässt sich auch als Widerstreit zwischen zwei fundamentalen normativen Anforderungen an den demokratischen öffentlichen Raum ausdeuten, wie sie jüngst erneut Jürgen Habermas herausgestellt hat. Wie in der Einleitung skizziert wurde, besteht das schädliche Potenzial von Fake News nicht nur darin, dass diese Einzelpersonen täuschen können. In ihrer Breitenwirkung besteht ein Teil des Schadens, den Fake News verursachen können, schließlich in ihrer Tendenz den Informationsraum zu verwirren und Meinungen so weit zu polarisieren, dass sinnvolle öffentliche Diskussionen erschwert werden. Damit aber wird eine zentrale Funktionsleistung von demokratischer Öffentlichkeit gefährdet, die Habermas darin sieht, Meinungen in gründegeleitete Diskurse einzubinden und zu rational akzeptablen Ergebnissen zu führen. Auf der anderen Seite steht nach Habermas jedoch die Forderung

24 Elizabeth Stewart: »Detecting Fake News. Two Problems for Content Moderation«, in: *Philosophy & Technology* 34 (2021), S. 923–940.

25 Diese Leitlinie besteht bei Facebook schon länger (siehe FN 20) und unabhängig von den (zur Zeit der Entstehung dieses Artikels) jüngeren Ankündigungen von Mark Zuckerberg, den Faktencheck zukünftig generell auf den zu Meta gehörenden Plattformen abzuschaffen und die Content-Moderation auf Community Notes abzustellen. Ob dieser Moderationszustand in der Praxis mit dem Digital Service Act der EU überhaupt vereinbar ist, scheint allerdings juristisch fragwürdig. Siehe die Einschätzungen in: »Zuckerbergs Pläne zur Moderation von Inhalten bei Meta«, in: *science media center germany,* 10.1.2025, https://www.sciencemediacenter.de/angebote/zuckerbergs-plaene-zur-moderation-von-inhalten-bei-meta-25004 (aufgerufen: 24.1.2025).

an die Öffentlichkeit inklusiv zu sein, das heißt, die öffentliche Artikulation verschiedener Meinungen, insbesondere von den Personen, die von kollektiv bindenden Entscheidungen betroffen sind, zu ermöglichen.[26] An den oben diskutierten Grenzfällen lässt sich nun besonders eindrücklich zeigen, dass sich die Content-Moderation als situiert in einer Spannung zwischen diesen beiden Anforderungen deuten lassen kann. Derart normativ umgedeutet, stellt sich auch für die Grenzfälle die Frage: Sollte man hier die Inklusion potenzieller Fake News im Grenzfallgebiet favorisieren, womöglich auf Kosten vernunft- und wahrheitsgeleiteter öffentlicher Diskurse? Oder sollte man primär den vernunftorientierten Charakter der Öffentlichkeit bewahren, womöglich unter Einschränkung seiner Inklusivität? In diesem grundlegenden Spannungsfeld lässt sich auch die Praxis der Content-Moderation bei der Frage nach der Einordnung als ›fake‹ oder ›nicht-fake‹ verstehen.[27]

Man mag hier einwenden wollen, dass dies womöglich arg zugespitzt ist und sich der Wertekonflikt nur dann ergibt, wenn man das *Löschen* von Nachrichten als primäre Aufgabe der Content-Moderation ansieht. Dann liegt es auch besonders nahe, freie Meinungsäußerung bzw. Inklusivität als einen zentralen Wertpol zu akzentuieren, und auch, wie Stewart es formuliert, ›Zensur‹ als eine mögliche Gefahr zu benennen.[28] Jedoch: Auch wenn man andere Arten der Intervention durch Content-Moderation betrachtet, etwa erstens die Reduktion von Sichtbarkeit oder zweitens das Versehen mit Warnhinweisen,[29] so zeigen sich diese im ähnlichen Sinne als

26 Habermas: *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*, S. 21.

27 Was nicht bedeutet, dass diese Spannung nicht auch weitere Implikationen hat, etwa auf dem Feld der journalistischen Aufbereitung von Themen. Neben dem Grundsatz, dass hierbei keine Fehlinformationen verbreitet werden sollten, ist dabei natürlich auch die Ausgewogenheit in der Darstellung entscheidend und kann als ein Gütekriterium verstanden werden, das versucht die Wahrheitsorientierung bei gleichzeitiger Inklusion möglichst diverser Sichtweisen zu vereinbaren.

28 Die zugespitzte Frage nach dem Löschen von Inhalten tangiert dabei auch besonders prägnant normative Vorstellungen darüber, in welchem Maße gewissermaßen paternalistisch in den öffentlichen Raum eingegriffen werden sollte bzw. in welchem Maße vielmehr die eigenständige Urteilsbildung der Rezipient:innen von Nachrichten hochzuhalten ist.

29 Auch dies sind bei Weitem nicht die einzigen Interventionsmöglichkeiten von Content-Moderation. So gibt es auch stärker rezipientenbezogene Eingriffsmöglichkeiten, die etwa darin bestehen, Falschaussagen durch Links zu entsprechenden Faktencheck-Webseiten zu entkräftigen oder Indikatoren hervorzuheben, die die Glaubwürdigkeit eines Medieninhalts in Frage stellen. Das damit verbundene Ziel kann hierbei

wertgeladen. Offenbar nämlich stellen sie eine werteabwägende Intervention in den öffentlichen Raum dar. Im ersten Fall etwa wird eine Verteilung der Ressource Aufmerksamkeit vorgenommen, die gedacht werden muss als von der evaluativen Frage geleitet, welche Meldung es *wert* ist *so und so viel* öffentliche Aufmerksamkeit zu erhalten. Im zweiten Fall wird darüber entschieden, welche Meldung es *wert* ist, gewissermaßen mit einer Kredibilitätshürde versehen zu werden. In beiden Fällen findet also eine Entscheidung darüber statt, wie *wertvoll* bzw. irrelevant eine Meldung *für die Öffentlichkeit* ist. Auch solche Entscheidungen lassen sich somit als an der Frage orientiert deuten, *welches Maß und welche Art von Inklusion* bestimmte Meldungen verdienen, auch angesichts ihrer potenziellen Gefahr, die Wahrheits- und Vernunftorientierung der Öffentlichkeit zu unterminieren.

Der Punkt bleibt also ganz grundlegend, dass solche Entscheidungen eine normativ geleitete Formierung der Öffentlichkeit vornehmen. Sie manifestieren insofern eine bestimmte Politik des öffentlichen Raumes. Und eine solche Politik des öffentlichen Raumes, so soll im Folgenden gezeigt werden, schreibt sich auch in die Algorithmen zur Detektion von Fake News fort. Zunächst aber werde ich kurz die Funktionsweise dieser Technik grob skizzieren.

Zur automatisierten Detektion von Fake News[30]

Das zentrale Ziel der Technik der automatisierten Fake-News-Detektion (FND) liegt darin, großflächiger, schneller und kostengünstiger als allein

auch darin bestehen, Rezipient:innen im Umgang mit Medieninhalten kritisch zu schulen. Für eine Übersicht über mögliche Interventionsmöglichkeiten und deren Prominenz in der Forschung: Katrin Hartwig, Frederic Doell und Christian Reuter: »The Landscape of User-centered Misinformation Interventions – A Systematic Literature Review«, in: *ACM Computing Surveys* 56 (2024), Heft 11, S. 1–36. Es scheint mir aber auch für diese die grundlegende Idee einschlägig, dass mit solchen Interventionen eine Entscheidung getroffen wird, die Auswirkungen auf die Kredibilität von zur Diskussion stehenden Medieninhalten hat und somit ebenfalls im Rahmen der angesprochenen normativen Spannung operiert.

30 Neben dem Forschungsbereich der automatisierten Fake-News-Detektion gibt es noch andere Bereiche der generellen automatisierten Erkennung von Desinformation. So gibt es beispielsweise eine eigene Forschungslinie der sogenannten Rumour Detection oder der Erkennung von Deep Fakes. Inwieweit die hier entfaltete Argumentation auch für diese einschlägig ist, wäre gesondert zu prüfen.

menschliche Faktenchecker:innen arbeiten zu können. Dies ist insbesondere in den sozialen Medien relevant, in denen Fehlinformationen schnelle Verbreitung finden und schnell irreversible Schäden verursachen können.[31] Ein strategischer Schwerpunkt, der mit FND verbunden ist, liegt deshalb darin, Fake News möglichst instantan zu identifizieren und zu entfernen bzw. einzudämmen.[32]

Dabei gibt es im Bereich der automatisierten FND eine schier unüberblickbare Reihe verschiedener Ansätze aus dem Bereich der KI bzw. des Maschinellen Lernens, speziell dem Natural Language Processing und der Graphentheorie.[33] Im Groben wird hier zwischen zwei Grundansätzen unterschieden, und zwar zwischen einerseits wissensbasierten und andererseits merkmalbasierten Ansätzen.[34] Bei ersteren wird auf eine Wissensbasis zurückgegriffen, anhand derer Aussagen von Meldungen direkt auf ihren Wahrheitsgehalt hin überprüft werden können. Zweitere versuchen demgegenüber anhand gewissermaßen indirekter Merkmale zu erkennen, ob es sich bei einer Nachrichtenmeldung um Fake News handelt. In Bezug auf diese Ansätze wird üblicherweise noch binnendifferenziert zwischen solchen Ansätzen, die Fake News an Merkmalen des Textes, an Merkmalen der Quelle oder anhand von Charakteristika ihres Verbreitungs- und Interaktionsmusters zu erkennen versuchen.[35]

In der folgenden Darstellung werde ich nicht alle Ansätze erläutern können und mich stattdessen auf die grundlegende Funktionsweise der

31 Shubhangio Rastogi und Divya Bansal: »Time is Important in Fake News Detection: A Short Survey«, in: *International Conference on Computational Science and Computational Intelligence (CSCI) 2021*, Las Vegas, NV, USA, S. 1441–1443.

32 Was nicht heißen soll, dass Fake News als Problem allein in der Sphäre der sozialen Medien bleiben. Dies gilt allein deshalb nicht, da sie auch im erweiterten öffentlichen Raum, etwa in den klassischen Medien aufgegriffen werden können. Siehe etwa: Giansiracusa: *How Algorithms Create and Prevent Fake News*, Kap. 1.

33 Siehe etwa die Übersicht in: Zhou u.a.: »A Survey of Fake News«: in: *ACM Computing Surveys* 53. Aktuell wird auch zunehmend an den Möglichkeiten zum Einsatz von *Large Language Models* geforscht, siehe etwa: Eleftheria Papageorgiou u.a.: »A Survey on the Use of Large Language Models (LLMs) in Fake News«, in: *Future Internet* 16 (2024), Heft 8, 298.

34 Vgl. Suhaib Kh Hamed u.a.: »A Review of Fake News Detection Approaches. A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion«, in: *Heliyon 9* (2023), Heft 10, e20382.

35 Ebd.