Katzenbach, Christian

Der „Algorithmic turn“ in der Plattform-Governance.

KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie

Köln Z Soziol (2022) (Suppl 1) 74:283–305
https://doi.org/10.1007/s11577-022-00837-4

Kölner Zeitschrift für Soziologie und Sozialpsychologie

ABHANDLUNGEN

Der „Algorithmic turn“ in der Plattform-Governance.

Die diskursive, politische und technische Positionierung von Algorithmen und KI als „technological fix“ für komplexe Herausforderungen

Christian Katzenbach

Eingegangen: 8. Dezember 2021 / Angenommen: 17. März 2022 / Online publiziert: 2. Mai 2022

Zusammenfassung Die Regulierung von Plattformen ist zu einem zentralen Thema öffentlicher und politischer Debatten geworden: Wie sollen einerseits Anbieter von sozialen Medien mit problematischen Inhalten wie Misinformation und Hassreden umgehen? Und wie sollten wir andererseits Plattformen regulieren, z. B. indem sie für Inhalte haftbar gemacht werden oder zum Einsatz von Upload-Filtern gedrängt werden? Vor diesem Hintergrund rekonstruiert der Beitrag einen „algorithmic turn“ in der Governance von Plattform, d. h. der zunehmenden Positionierung von automatisierten Verfahren zur Adressierung dieser Governance-Fragen. Dabei arbeitet der Beitrag heraus, dass dies eine Entwicklung ist, die keineswegs nur durch technische Fortschritte in der Klassifikation von Inhalten zu erklären ist. Automatisierte Verfahren können nur als schlüssige Lösung für komplexe Verfahren positioniert werden, weil sie sich günstig in diskursive und politische Entwicklungen einbetten lassen. Der Beitrag identifiziert einen diskursiven „responsibility turn“ der zunehmenden Zuschreibung von Verantwortung an die Plattform und eine politisch-regulative Entwicklung der zunehmenden Mithaftung von Plattformen für Inhalte. Dafür kombiniert der vorliegende Beitrag techniksoziologische und institutionentheoretische Perspektiven. Im Schlussabschnitt werden die breiteren Entwicklungslinien einer zunehmenden Institutionalisierung und „Infrastrukturisierung“ von algorithmischen Systemen reflektiert. Der Beitrag identifiziert unter anderem die Gefahr, dass mit der Verlagerung von Entscheidungen über umstrittene Inhalte in Technik und Infrastruktur diese inhärent politischen Fragen der öffentlichen Debatte entzogen und der Entscheidungshoheit der Plattformen überlassen werden.

C. Katzenbach
Zentrum für Medien-, Kommunikations- und Informationsforschung (ZeMKI), Universität Bremen
Linzer Straße 4, 28359 Bremen, Deutschland

Alexander von Humboldt Institut für Internet und Gesellschaft
Franz. Straße 9, 10117 Berlin, Deutschland
E-Mail: katzenbach@uni-bremen.de

Schlüsselwörter Social Media · Regulierung · Techniksoziologie · Diskurse · Plattformen

The Algorithmic Turn in Platform Governance

The Discursive, Political, and Technological Positioning of Algorithms and AI as a "Technological Fix" for Complex Challenges

Abstract The governance of platforms has become a central topic of public and political debate: How should social media providers deal with problematic content such as misinformation and hate speech? And how should we regulate platforms—make them liable for content or push them to use upload filters, for example? Against this background, the article reconstructs the current "algorithmic turn" in platform governance as a development that can by no means be explained only by technical advances in content classification. Automated procedures can be positioned as a conclusive solution for complex procedures only because they can be favourably embedded in discursive and political developments—above all, as a discursive "responsibility turn" of increasing attribution of responsibility to platforms and as a political–regulatory development of increasing co-liability of platforms for content. This paper reconstructs and discusses these issues from a conceptual perspective that combines work in the sociology of technology and institutional theory. The final section reflects on the broader lines of an increasing institutionalisation and "infrastructuralisation" of algorithmic systems. Among other things, the paper identifies the danger of, by shifting decisions about contentious content to technology and infrastructure, removing these inherently political questions from public debate and leaving them to the decision-making authority of platforms.

Keywords Social media · Regulation · Science and technology studies (STS) · Discourses · Platforms

1 Einleitung

Die Governance und Regulierung von Plattformen ist zu einem zentralen Thema öffentlicher und politischer Debatten geworden: Wie sollen Anbieter von sozialen Medien mit problematischen Inhalten wie Misinformation und Hassreden umgehen? Können und sollen sie selbst Unterscheidungen treffen und durchsetzen zwischen wahren und unwahren Behauptungen, zwischen legitimen Äußerungen und grenzüberschreitender Hassrede? Gleichzeitig wird im politischen Feld intensiv diskutiert, wie Plattformen selbst reguliert werden sollten, um mögliche negative Auswirkungen auf Demokratie und gesellschaftlichen Zusammenhalt zu reduzieren. Sollten sie etwa für Inhalte haftbar gemacht werden oder zum Einsatz von Upload-Filtern gedrängt werden? Während sich inzwischen ein gesellschaftlicher Konsens etabliert hat, dass Plattformen stärker als bislang Verantwortung für ihre Inhalte übernehmen müssen, besteht wenig Einigkeit und Klarheit darüber, in welcher Form dieser Konsens praktisch umgesetzt werden sollte.

In der Praxis des täglichen Betriebs der größeren Plattformen hat sich eine (teil-)automatische Bearbeitung dieser Fragen längst etabliert: Der flächendeckende Einsatz von algorithmischen Systemen der Inhalteklassifikation und -filterung ist inzwischen Standard (Gorwa et al. 2020). Mit Blick auf urheberrechtliche Fragen setzen YouTube und andere Musik- und Videoplattformen schon seit mehr als einer Dekade automatisierte Systeme zur Erkennung und Regulierung von Inhalten ein. In den vergangenen Jahren werden solche Verfahren zunehmend auch zur Adressierung von Hassrede, Misinformation, terroristischen und anderen gewaltvollen Inhalten eingesetzt. Wenn Nutzerinnen und Nutzer Inhalte hochladen, werden sie schon beim Upload mit industrieweiten Datenbanken wie etwa PhotoDNA abgeglichen, die bereits identifizierte Inhalte mit Abbildungen von Kindesmissbrauch gespeichert haben. Zur Reduktion von Hassrede setzen die meisten Plattformen inzwischen Verfahren ein, die Texte anhand ihrer vermeintlichen „toxicity“ bewerten und ab einem bestimmten Niveau sperren.

Dieser „algorithmic turn“ in der Governance und Regulierung von Plattformen lässt sich aber nicht allein durch technische Entwicklungen und Fortschritte erklären. Er ist vielmehr Resultat einer Verknüpfung von politischen, diskursiven und technischen Bewegungen, die sich wechselseitig verstärken – so die Leitthese des vorliegenden Beitrags. Fraglos profitieren solche Systeme von technischen Fortschritten in der Erkennung und Klassifikation von Inhalten. Gleichzeitig werden Plattformen aber zunehmend stärker in die auch rechtliche Verantwortung für ihre Inhalte genommen und zum Monitoring problematischer Inhalte verpflichtet. Zudem fügt sich die Positionierung von automatisierten Systemen als adäquate Lösung für strittige Fragen wie Hassrede und Misinformation in die lange Tradition des „technological fix“ ein – der Positionierung von Technik als vermeintlicher Lösung für komplexe, gesellschaftliche Fragen.

Vor diesem Hintergrund rekonstruiert der vorliegende Beitrag den „algorithmic turn“ in der Governance von und durch Plattformen. Dafür wird auf techniksoziologische und institutionentheoretische Konzepte zurückgegriffen, die in den ersten beiden Abschnitten eingeführt werden: den „technological fix“ als wiederkehrendes Motiv im Verhältnis von Technik und Gesellschaft sowie eine institutionentheoretische Perspektive auf Governance und Algorithmen. Dann werden sukzessive diskursive, politisch-regulatorische und technische Entwicklungen rekonstruiert, die gemeinsam einen „algorithmic turn“ ermöglichen und so Algorithmen und künstliche Intelligenz (KI) als adäquate Antworten auf komplexe Herausforderungen in der Moderation und Regulierung positionieren. Im Schlussabschnitt werden die breiteren Entwicklungslinien einer zunehmenden Institutionalisierung und „Infrastrukturisierung“ von algorithmischen Systemen reflektiert. Der Beitrag identifiziert unter anderem die Gefahr, dass mit der Verlagerung von Entscheidungen über umstrittene Inhalte in Technik und Infrastruktur diese inhärent politischen Fragen der öffentlichen Debatte entzogen und der Entscheidungshoheit der Plattformen überlassen werden.

2 Das Motiv des „Technological fix“

Im Kontext der intensiven Debatte um Facebooks Rolle in der US-Wahl 2016 und in tiefgreifende Eingriffe in die Privatsphäre und den Datenschutz von Millionen von Nutzern (Cambridge Analytica) wurde Facebook CEO Marc Zuckerberg im April 2018 zu einer gemeinsamen Anhörung des Rechts- und des Wirtschaftsausschusses des US-Senats geladen. Dabei befragten die Senatorinnen und Senatoren Zuckerberg nicht nur zu den Geschehnissen der Vorjahre, sondern auch nach den Plänen des Konzerns, zukünftig adäquat und verantwortungsvoll auf die Herausforderungen durch Desinformationskampagnen zu reagieren, aber auch auf die Verbreitung von Hassrede, terroristischen und anderen problematischen Inhalten. Zuckerbergs Antworten auf diese vielfältigen Fragen lassen sich recht einfach in einem kurzen Ausdruck zusammenfassen: „AI will fix this!“ (Katzenbach 2019). In unterschiedlichen Formulierungen verwies Zuckerberg immer wieder auf die Entwicklung und den zunehmenden Einsatz von KI-gestützten Systemen zur Erkennung von Hassrede, Terrorismus und Misinformation: „In the future, we’re going to have tools that are going to be able to identify more types of bad content.“[1] Auch die oft schwierige kontextabhängige und nuancenreiche Einordnung von Sprache werde zukünftig durch automatisierte Systeme zu bewältigen sein: „over a 5 to 10-year period, we will have A.I. tools that can get into some of the nuances – the linguistic nuances of different types of content to be more accurate in flagging things for our systems. But, today, we’re just not there on that.“ Egal welche Herausforderung, Zuckerberg positionierte neue Technik als Antwort auf komplexe soziale Herausforderungen.

Mit dieser Reaktion griff Zuckerberg auf ein regelmäßiges Motiv im Verhältnis von Technik und Gesellschaft zurück, das vor allem Akteure aus Wirtschaft und Technikentwicklung immer wieder aufs Neue machtvoll in Diskurse einbringen: den „technological fix“, die Positionierung von Technik als notwendige und funktionale Lösung sozialer Probleme und Herausforderungen. Rudi Volti hat in den 1990er-Jahren auf die Allgegenwart dieses Motivs hingewiesen:

> The list of technologies that have been or could be applied to the alleviation of social problems is an extensive one, and examples could be supplied almost indefinitely. What they have in common is that they are “technological fixes”, for they seek to use the power of technology in order to solve problems that are nontechnical in nature (Volti 2014, S. 30).

Ein blindes Vertrauen auf die Wirkung der Technik sei dabei in der Regel mit einer Ignoranz gegenüber der Gründe und Dynamik der bestehenden sozialen Probleme verbunden: „In dieser Sichtweise bewältigen Verkehrsleitsysteme das steigende Pkw-Aufkommen in den Städten (und der Autoverkehr wird nicht reduziert), Nahrungsmittelimporte bewahren die Ärmsten vor dem Verhungern (und die Ursachen

[1] Vergleiche das Video und die offizielle Dokumentation des Hearings auf der Website des US-Senats (https://www.judiciary.senate.gov/meetings/facebook-social-media-privacy-and-the-use-and-abuse-of-data) sowie das wörtliche Transkript der Befragung unter: https://en.wikisource.org/wiki/Zuckerberg_Senate_Transcript_2018.

werden nicht bekämpft), Rinder werden gekeult (und nicht die industrielle Massentierhaltung verabschiedet)“ (Degele 2002, S. 25).

Im Kontext der Digitalisierung ist Zuckerberg keinesfalls der Einzige, der diese Erzählung weiterträgt – es ist vielmehr ein zentrales Motiv der narrativen Grundzüge der US-dominierten Digitalbranche (Daub 2020). Evgeny Morozov etwa beschreibt eindrücklich, wie Silicon-Valley-Unternehmen soziale Probleme oder komplexe Zusammenhänge wie Gesundheit und Mobilität als eindeutig funktional lösbare Probleme behandeln. Durch die Bereitstellung immer neuer Dienste und Apps versprechen sie eine Optimierung der Prozesse. Was bei Volti „technological fix“ heißt, nennt Morozow „solutionism“: „Recasting all complex social situations either as neat problems with definite, computable solutions or as transparent and self-evident processes that can be easily optimized – if only the right algorithms are in place!“ (Morozov 2013, S. 5).

Die gegenwärtige Hinwendung zur Adressierung komplexer Fragen der Gestaltung und Ordnung öffentlicher Kommunikation durch Algorithmen und künstliche Intelligenz ist also keineswegs einzigartig, sondern reiht sich ein in ein langwährendes Motiv im Verhältnis von Gesellschaft und Technik. Für die Einordnung des aktuellen „algorithmic turn“ ist dieser Befund wichtig, da er schon darauf hinweist, dass sich diese Hinwendung nicht allein durch etwaigen technischen Fortschritt erklären lässt. Es ist vielmehr ein immer wiederkehrendes Motiv, das je nach gesellschaftlicher Problemlage an Bedeutung gewinnt oder verliert. Der Aufwind für die prominente Positionierung eines „technological fix“ als machtvolles Deutungs- und Erklärungsmuster kann durch technische Impulse angestoßen werden, keinesfalls aber allein getragen werden.

3 Institutionentheoretische Perspektive auf Governance und Algorithmen

Um der Frage nachzugehen, wie der gegenwärtige „algorithmic turn“ an Bedeutungskraft gewonnen hat im Zusammenspiel verschiedener Kräfte und Dimensionen, nutzt der vorliegende Beitrag eine institutionentheoretische Perspektive auf Governance und Algorithmen, die Schwerpunkte legt auf (a) den *Prozess* der Formation institutioneller Zusammenhänge und (b) auf das *Wechselverhältnis* verschiedener Dimensionen institutioneller Zusammenhänge.

In den vergangenen Jahren sind Algorithmen, aber auch allgemeiner Medien und Governance vermehrt unter institutionentheoretischen Perspektiven beschrieben worden. Napoli (2013) und Katzenbach (2012) haben früh Algorithmen und allgemein Medientechnologien als Institutionen beschrieben. Gerade im Kontext des aktuellen, tiefgreifenden Medienwandels greifen Kommunikations- und Medienwissenschaftler vermehrt auf Institutionentheorien zurück (Jarren 2019; Sandhu 2018; Katzenbach 2017; Künzler et al. 2013). Dabei wird der Institutionenbegriff mal stärker mit Bezug auf Traditionslinien des Historischen Institutionalismus und des ökonomischen Neoinstitutionalismus als Organisationen, Gesetze und andere formale Regelwerke verstanden; zunehmend werden aber Bezüge auf soziologische Varianten des Neoinstitutionalismus genutzt, die auch Erwartungen und Diskurse in

den Institutionenbegriff hineinholen (vgl. Hall und Taylor 1996; DiMaggio 1998; als Überblick Katzenbach 2017, S. 88–96).

Im Anschluss an die soziologischen Institutionentheorien werden im vorliegenden Beitrag Institutionen allgemein als Regelungs- und Erwartungsstrukturen verstanden, die legitimes Handeln und Entscheiden begründen (Hasse und Krücken 2005, S. 7). Dadurch lassen sich komplexe Fragen sowohl der Plattform-Governance als auch der Rolle von Algorithmen adäquat als ein vielschichtiges Phänomen beschreiben, in dem verschiedene Faktoren und Dimensionen zusammenwirken. Für dessen Strukturierung hat es sich als hilfreich erwiesen, an Scotts Dimensionierung von Institutionen anzuknüpfen (Scott 2008) und so die durch Institutionen geleistete Orientierungs- und Ordnungsfunktion analytisch auf drei Ebenen zu betrachten (Donges 2006; Puppis 2010; Katzenbach 2017):

- auf der *regulativen Ebene* der Etablierung und Durchsetzung formaler Regeln;
- auf der *diskursiven* Ebene, der Konstruktion, Deutung und Wahrnehmung von sozialer Wirklichkeit und damit verbundenen Handlungsoptionen; sowie
- einer *technologischen Ebene*, auf der soziale Erwartungen und Regeln in technisch-materiellen Artefakten und Strukturen verkörpert, übersetzt und weiterentwickelt werden.

Eine solche integrierte Perspektive macht deutlich, dass sich Ordnungs- und Wandlungsprozesse, etwa zu Regeln und Strukturen der medialen Kommunikation, wie im Bereich Plattform-Governance, über ganz unterschiedliche Mechanismen und an verschiedenen Orten realisieren. Es sind nicht allein Technologien, die Wandel anstoßen, oder allein formale Regeln, die soziale Zusammenhänge ordnen – auch Diskurse, entstehende Sichtweisen und Deutungsmuster erzeugen Wandel und stellen kollektive Verbindlichkeit her.

Auf dieser institutionentheoretischen Basis lässt sich nun strukturiert herausarbeiten, dass Algorithmen und KI sich nicht einfach als funktionale Lösung eines sozialen Problems – z. B. Hassrede auf Plattformen – durchsetzen, sondern vielmehr in einem mehrschichtigen Institutionalisierungsprozess als adäquat erscheinende Lösung für die komplexen Herausforderungen im Bereich Plattform-Governance erfolgreich positioniert werden. Dazu werden nun im Folgenden die Entwicklungen in den drei institutionentheoretisch begründeten Dimensionen rekonstruiert.

4 Die Diskursive Wende: Vom „Responsibility turn“ zum „Technological fix“

Auf der diskursiven Ebene lässt sich mit Blick auf Plattform-Governance eine Wechselbewegung identifizieren: Öffentlichkeit und Politik fordern von Plattformen in zunehmendem Maße die Übernahme von Verantwortung für die Inhalte und Kommunikationsdynamiken auf ihren Diensten („responsibility turn“); Plattform-Betreiber reagieren wiederum auf diesen wachsenden Druck mit dem Versprechen einer technischen Lösung.

Plattformunternehmen, wie Facebook, Twitter und Google, aber auch Uber und Airbnb, waren lange sehr erfolgreich darin, sich selbst als neutrale Vermittler zu posi-

tionieren (Gillespie 2010). Die absolute Ermöglichung freier Meinungsäußerung und das „Information-wants-to-be-free“-Mantra waren lange Zeit zentrale Leitbilder der Silicon-Valley-Unternehmen (Vaidhyanathan 2012). Ob Suchergebnisse, Newsfeed oder einfach die Bereitstellung der Möglichkeit, sich öffentlich zu äußern: Bis weit in die 2010er-Jahre hinein haben Plattformunternehmen ihre Dienste als wertfreie Angebote und sich selbst als IT-Unternehmen und nicht als Medienorganisationen dargestellt (Napoli und Caplan 2017). Die eingesetzten Algorithmen zur Sortierung und Priorisierung von Inhalten würden objektive und damit neutrale Resultate liefern (Ames 2018). Diese „spiritual deferral to algorithmic neutrality“ (Morozov 2011, o. S.) drückte sich etwa in Googles systematischer Weigerung aus, für Suchergebnislisten Verantwortung zu übernehmen und sie manuell zu ändern, auch wenn sie rassistische oder diskriminierende Inhalte auf den ersten Plätzen zeigten (Gillespie 2014, S. 180–181; Noble 2018). Twitter machte bis 2015 gleich in den ersten Worten seiner „Twitter Rules“, gleichzeitig Selbstdarstellung und Regelwerk, deutlich, dass die Nutzer für ihre Inhalte selbst verantwortlich sind und Twitter als Anbieter neutral bleibe: „We respect the ownership of the content that users share and each user is responsible for the content he or she provides. Because of these principles, we do not actively monitor and will not censor user content.“[2] Diese Positionierung als neutrale Vermittler war für die Plattformanbieter hochgradig attraktiv, da sie sich so als selbstverständliches und bald scheinbar unverzichtbares Element alltäglicher Kommunikation etablieren konnten, gleichzeitig aber weder sozial verantwortlich noch rechtlich haftbar für die Inhalte zu machen waren. „They do so strategically, to position themselves both to pursue current and future profits, to strike a regulatory sweet spot between legislative protections that benefit them and obligations that do not, and to lay out a cultural imaginary within which their service makes sense“ (Gillespie 2010, S. 348). Diese Positionierung begünstigte ganz wesentlich das rasante Wachstum der Plattformen zu zentralen Institutionen gesellschaftlicher Kommunikation (Eisenegger 2021).

Doch spätestens seit den US-Wahlen 2016 und den zunehmenden politischen und gesellschaftlichen Konflikten in Migrationsfragen seit 2015 lässt sich ein „responsibility turn“ in der Debatte um Plattformen beobachten. In diesen Jahren haben die Kontroversen um die Rolle und die adäquate Regulierung von Plattformen deutlich zugenommen (Katzenbach 2021). Besonders die intensiven Debatten um Misinformation oder Fake News (Righetti 2021) und Hatespeech (Tworek 2021) haben Fragen der Macht und der Verantwortung von Plattformen weit oben auf die öffentlichen und politischen Agenden platziert. Die Anbieter werden im Diskurs nun zuvorderst als aktive Akteure wahrgenommen, die ihre eigenen Dienste auf spezifische Weisen organisieren und dabei sowohl eigene Interessen verfolgen (quantitative Optimierung auf Interaktionen, Monetarisierung durch Werbung) als auch externe Effekte (Verstärkung von Dynamiken der Misinformation, Hasskommentare) erzeugen. An diesen „responsibility turn“ haben sich Plattformanbieter inzwischen merklich angepasst, indem sie Fehler eingestehen, Verantwortung annehmen und – etwa Facebook – in der Außenkommunikation ihre Mission der Vernetzung alle Menschen

[2] „The Twitter Rules“, ca. 2009–2015. Archiviert unter: https://web.archive.org/web/20090118211301/http://twitter.zendesk.com/forums/26257/entries/18311 (Zugegriffen: 22. Mai 2021).

stärker qualitativ als quantitativ ausdeuten (Lischka 2019; Haupt 2021). Seit 2018 beginnt etwa Twitter seine Regeln mit den Worten: „Twitter's purpose is to serve the public conversation. Violence, harassment and other similar types of behavior discourage people from expressing themselves, and ultimately diminish the value of global public conversation. Our rules are to ensure all people can participate in the public conversation freely and safely." Im starken Gegensatz zur ursprünglichen Formulierung zeigt sich hier deutlich der „responsibility turn": Twitter übernimmt Verantwortung für die Inhalte auf seinem Dienst und begründet Einschränkungen der Meinungsfreiheit.

Während also inzwischen weitgehend ein Konsens darüber herrscht, dass Plattformen Verantwortung für die Inhalte und Kommunikationsdynamiken auf ihren Diensten innehaben, ist die Ausgestaltung dieser Verantwortung keineswegs geklärt. Dabei stellt sich nicht nur die schon komplexe Frage, wie und entlang welcher Kriterien Plattformen kontroverse und problematische Inhalte beurteilen und gegebenenfalls sperren sollten. Sondern es besteht eine besondere Herausforderung darin, diese oft schwierigen Abwägungen zwischen Informations- und Meinungsfreiheit einerseits und etwa Persönlichkeitsrechten andererseits „at scale" durchzuführen, d. h. in der massenhaften Anzahl von Inhalten auf Plattformen.

In dieser Situation, massenhaft potenziell problematische Inhalte einerseits und wachsende Zuschreibung von Verantwortung von Plattformen andererseits, erscheint nun für viele der „technological fix" als der einzige Ausweg. Anbieter und Regulierungsakteure haben sich in den vergangenen Jahren wechselseitig darin bestärkt, dass Technologie, insbesondere „künstliche Intelligenz" (KI), die Probleme sozialer Medien lösen können: „AI will fix this!", verspricht Facebooks CEO Zuckerberg regelmäßig in Anhörungen in Nordamerika und Europa (Katzenbach 2019; Lischka 2019; Russell 2019). Yann LeCun, weltweit führender KI-Forscher in Diensten Facebooks, sieht Fake News oder Misinformation als technisch lösbar an (Seetharaman 2016) und Journalisten folgen ihm regelmäßig in dieser Sichtweise (vgl. z. B. Elgan 2017). In der neuen EU-Urheberrechtslinie und wie im EuGH-Urteil zu Facebooks Löschpflichten werden funktionierende automatische Filtersysteme offenbar vorausgesetzt, ohne ihre Beschränkungen und ihre Nebeneffekte auf Meinungsfreiheit substanziell zu diskutieren (vgl. Abschn. 6).

Damit knüpft die öffentliche Debatte zur Governance von und durch Plattformen an das allgemeine Motiv des „technological fix" an. Gleichzeitig ist die Debatte eingebunden in die breitere aktuelle Positionierung von KI als Problemlöser. Etwa zeitgleich mit dem „responsibility turn" ist die Aufmerksamkeit für KI in der öffentlichen und politischen Debatte massiv gestiegen. Während zwar zunehmend auch gesellschaftliche Probleme und Herausforderungen thematisiert werden, so dominieren doch Produkte und vermeintliche Innovationen die Berichterstattung deutlich (Brennen et al. 2018; Puschmann und Fischer 2021). Gerade im Fall von KI und den oft kaum rekonstruierbaren Entscheidungsgrundlagen werden Technologien magische Eigenschaften zugeschrieben (Bory 2019, Cave und Dihal 2019). Dieser „enchanted determinism" (Campolo und Crawford 2020) sorgt hier für eine besonders starke Ausprägung des „Technological-fix"-Motivs. Die jüngere Debatte um den „technological fix" für die komplexen Herausforderung in der Moderation und Regulierung von Social-Media-Inhalten kann hier anknüpfen. Dieser diskursive

Resonanzboden erlaubt es Plattformen erst, auf vermeintlich überzeugende Weise, technische Lösungen als adäquate Reaktion auf die wachsende Verantwortung für Inhalte in Position zu bringen.

5 Die politisch-regulative Wende: Vom Haftungsprivileg zu proaktiven Maßnahmen

Auf der politisch-regulatorischen Ebene hat besonders im Europäischen Kontext, aber auch in den USA, eine Suchbewegung eingesetzt, wie sich dieser „responsibility turn“ konkret in Recht übersetzen lässt. Im Ergebnis werden die Plattformanbieter in Gerichtsurteilen höchster Instanz sowie in zahlreichen regulatorischen Maßnahmen zunehmend stärker in die Pflicht und teilweise auch in die Haftung genommen.

Aus rechtlicher Sicht sind Plattformen und Anbieter sozialer Medien zunächst natürlich Gegenstand allgemeiner Gesetze sowie zahlreicher sektorspezifischer Regelungen. Prägend für die Regulierung von Plattformen der vergangenen Jahrzehnte ist das sogenannte Haftungsprivileg oder das „Notice-and-Takedown“-Verfahren: Erst wenn Anbieter Kenntnis von illegitimen Inhalten oder rechtswidrigen Haltungen auf ihren Diensten haben, müssen sie tätig werden, indem sie Inhalte filtern oder den Zugang von Nutzern sperren. Dieses Paradigma ist sowohl in Europa mit der E-Commerce-Richtlinie der EU 2001 (Kuczerawy und Ausloos 2015) als auch in den USA im US Digital Millennium Copyright Act (DMCA) und mit noch weitergehenden Freiheiten in Sektion 230 des US Communication Decency Act (CDA) seit etwa der Jahrtausendwende festgeschrieben (Citron und Wittes 2017; Gasser und Schulz 2015; Holland et al. 2016).

In Deutschland hat sich diese nachgelagerte Verantwortung von Anbietern sozialer Medien etwa im Telemediengesetz (TMG) ausgedrückt. Anbieter von Telemedien, darunter Plattformen, sind für bereitgestellte „eigene Informationen“ rechtlich verantwortlich (§ 7). Da die Funktion von sozialen Medien in der Rechtsprechung in der Regel so verstanden wird, dass sie – in den Worten des TMG – nur „fremde“ Informationen „übermitteln oder … den Zugang zur Nutzung vermitteln“, fallen die strittigen Inhalte nicht unter § 7. Für die vermittelten Inhalte aber sind sie nach § 8 TMG nicht verantwortlich und auch nicht haftbar zu machen. Erst wenn sie Kenntnis über den illegalen Inhalt erhalten, müssen sie eingreifen (§ 10).

Seit knapp 10 Jahren zeigt sich allerdings eine Entwicklung hin zu einer deutlich engeren Auslegung dieses Haftungsprivilegs bis hin zur Abkehr von diesem Paradigma. Diese „road to responsibilities“ (Síthigh 2020) drückt sich sowohl in Gerichtsurteilen als auch in regulatorischen Initiativen aus. In Deutschland hatte etwa der Bundesgerichtshof (BGH) in einem zentralen Urteil 2012 noch das Haftungsprivileg von Anbietern betont.[3] Im Jahr 2013 reformulierte der BGH in einem Urteil zum File Hoster Rapidshare das Haftungsprivileg dann schon deutlich voraussetzungsvoller, indem er dem Anbieter eine „Marktbeobachtungspflicht“ auferlegte, die verpflichtet, „mit geeignet formulierten Suchanfragen und … unter Einsatz von

[3] BGH, 27.03.2012, Az. VI ZR 144/11, URL: http://openjur.de/u/405723.html (Zugegriffen: 31. Mai 2021).

Webcrawlern zu ermitteln, ob sich hinsichtlich der konkret zu überprüfenden Werke Hinweise auf weitere rechtsverletzende Links auf ihrem Dienst finden."[4] In der Gesetzgebung in Deutschland kulminierte die Entwicklung der immer stärkeren Mithaftung von Anbietern im Netzwerkdurchsetzungsgesetz (NetzDG), das vor allem als Reaktion auf verstärkte Hasskommentare und Misinformation in sozialen Medien die großen Plattformen verpflichtet, in kurzen Zeitfenstern (24 Stunden bei „offensichtlich rechtswidrigen" Inhalten, 7 Tage bei allen weiteren) auf Beschwerden von Nutzern zu reagieren und Inhalte gegebenenfalls zu löschen (Schulz 2018).

Der Ende 2019 von den Bundesländern neu geschlossene Staatsvertrag zur Modernisierung der Medienordnung in Deutschland („Medienstaatsvertrag") geht ebenfalls in diese Richtung: Anbieter sozialer Medien werden nach Verabschiedung dieses Staatsvertrags durch die Länderparlamente als „Medienintermediäre" (im neuen § 2 Abs. 2 Nr. 16) in das bundesdeutsche System der Rundfunkmedienregulierung integriert. In der Folge werden sie die „zentralen Kriterien einer Aggregation, Selektion und Präsentation von Inhalten und ihre Gewichtung" (§ 93 Transparenz) offenlegen müssen. Zudem müssen sie sicherstellen, dass einzelne Inhalteanbieter nicht „diskriminiert" werden, d. h. die deklarierten Kriterien müssen unterschiedslos für alle Inhalte angewandt werden (§ 94 Diskrimierungsfreiheit). Die Auswirkungen dieser neuen Regulierungsformen müssen sich allerdings noch zeigen. Während die Diskriminierungsfreiheit eine regulatorische Novität darstellt, sind die Transparenzgebote und die verstärkte Mithaftung keine deutsche Besonderheit. In anderen Ländern, wie Frankreich und England, gehen Gesetzgeber und Regulierung ebenfalls in diese Richtung.

Auch Institutionen auf europäischer Ebene haben sich in Gerichtsurteilen höchster Instanz und Gesetzgebungsinitiativen vom Paradigma des Haftungsprivilegs schrittweise abgewendet und stärkere Anforderungen und Haftungsregeln eingeführt. Der Europäische Gerichtshof (EuGH) hat in Urteilen wie dem zum „Recht auf Vergessen" die Verantwortung der Plattformen für die von ihnen bereitgestellten Inhalte deutlich erhöht (Kuczerawy und Ausloos 2015) und verpflichtet zudem – etwa im Urteil Glawischnig-Piesczek vs. Facebook (EuGH, C-18/8) – Anbieter zum Einsatz automatisierter Verfahren, um – sind bestimmte Inhalte einmal als rechtswidrig notifiziert – die Veröffentlichung aller inhaltsgleichen Aussagen zu verhindern. EU-Kommission und Europäisches Parlament erhöhen durch die Einführung von Selbstverpflichtungen der Anbieter zur umgehenden Löschung und Blockierung von terror- und gewaltverherrlichenden Inhalten, durch die Verabschiedung der neuen Urheberrechtslinie mit deutlich weitergehenden Haftungsbestimmungen (Richtlinie EU 2019/790) die juristische Mitverantwortung der Anbieter sozialer Medien deutlich.

Der sich derzeit in der finalen Abstimmung der EU-Institutionen befindliche „Digital Services Act" (DSA)[5] setzt ebenfalls am Anspruch an, Plattformbetreiber deutlich mehr als bislang in die Verantwortung zu nehmen. Diese geplante Regu-

[4] BGH, 15.08.2013, Az. I ZR 80/12, Absatz 67. URL: http://lexetius.com/2013,3114 (Zugegriffen: 31. Mai 2021).

[5] Stand: Oktober 2021. Entwurfsfassung der EU-Kommission vom 15.12.2020 unter https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_de (Zugegriffen: 28. Okt. 2021).

lierung priorisiert allerdings deutlich stärker Maßnahmen der Rechenschafts- und Transparenzpflichten als Haftungsfragen. Das Haftungsprivileg soll im Grundsatz explizit nicht verändert werden, die Anforderungen werden aber entsprechend der o. g. Gerichtsurteile eng gefasst (DSA-Entwurf, Abschn. 2). Das zentrale Anliegen des DSA ist es aber, systematisch die von besonders großen Plattformen (mit mehr als 45 Mio. Nutzern monatlich, gemäß Art. 24) ausgehenden gesellschaftlichen Risiken (Misinformation, Hassrede, illegale Inhalte) zu adressieren. Dazu sollen diese Plattformen über (a) Transparenzregeln verpflichtet werden, ihre Funktionsweisen und Abwägungen in der Kuration von Inhalten „leicht verständlich“ öffentlich zu erklären (Art. 29, 1) sowie Nutzern die Option zur Änderung und Auswahl von Relevanzkriterien anzubieten, insbesondere die Möglichkeit, den Dienst ohne Profiling zu nutzen (Art. 29, 1–2). Gleichzeitig werden Plattformen bei Umsetzung des DSA (b) in vielfacher Weise stärker rechenschaftspflichtig: die Regulierung verlangt eine jährliche interne Risikobewertung (Art. 26), denen verpflichtend systematische Maßnahmen der Minderung von Risiken folgen müssen (Art. 27); zusätzlich werden diese Plattformen sich auf eigene Kosten einer externen Prüfung ihrer Pflichten und Verhaltenszusagen unterziehen (Art. 28) und Daten zur Evaluierung der Einhaltung dieser Regeln an eine nationale Regulierungsbehörde sowie der von ihnen ausgehenden Risiken an Forscher zur Verfügung stellen müssen (Art. 29).

Diese verschiedenen Maßnahmen auf der politisch-regulativen Ebene lassen sich zusammenfassen als eine Suchbewegung, die darauf zielt, die konsensuelle Forderung nach mehr Verantwortung der Plattformen in adäquate regulatorische Maßnahmen zu übersetzen. Deutlich zu beobachten ist eine sukzessive Abkehr vom Paradigma des Haftungsprivilegs. Der Zielpunkt ist derzeit noch nicht abzusehen; ein Wandel zum absoluten Gegenpol, der Haftung von Plattformen für alle Inhalte, die sie vermitteln, erscheint weder erstrebenswert noch realistisch – auch schlägt die EU mit dem DSA einen anderen Weg ein. Die Eingriffe in Meinungs- und Informationsfreiheit und der Verlust an vielfältigen Öffentlichkeiten wären ohnehin zu gravierend (Schulz 2019). Deutlich abzusehen ist aber bereits die merklich engere und voraussetzungsreichere Ausgestaltung von Haftungsausschlüssen, die mindestens proaktive Maßnahmen von Anbietern verlangen, um bereits bekannte illegale oder rechtsverletzende Inhalte zu blocken. Hinzu kommen die zunehmenden Ansprüche, auch schwierig zu klassifizierende Fragen wie Urheberrechtsverletzungen, Hassrede oder Misinformation effektiv zu erkennen und zu bekämpfen (Bloch-Wehba 2020).

Diese wachsenden politisch-regulativen Anforderungen an Plattformen begünstigen die hier dargestellte algorithmische Wende massiv. Denn wie lässt sich die Kategorisierung und gegebenenfalls Filterung von Inhalten „at scale“, d. h. in massenhafter Form, bewerkstelligen? Angesichts der Masse an Inhalten und der Größe der Probleme scheinen allein automatisierte Verfahren helfen zu können. Auf diese Weise forciert die gegenwärtige politisch-regulative Entwicklung den „algorithmic turn“ in der Governance von Plattformen.

6 Die Technische Wende: Von schwarzen Listen zu Klassifikationen

Diese diskursive und politisch-regulatorische Entwicklung ist natürlich nicht völlig unabhängig von der technischen Entwicklung. Auch wenn es ein Anliegen dieses Beitrags ist, darzustellen, dass der „algorithmic turn" nicht allein durch technischen Fortschritt zu erklären ist, so tragen technische Entwicklungen durchaus zu seiner Erklärung bei. In den vergangenen 10 Jahren gab es fraglos nicht nur einen KI-Hype, sondern auch massive Fortschritte in der automatisierten Erkennung von Mustern; im Aufbau von riesigen Datensets, um Prozesse des maschinellen Lernens durchzuführen und zu optimieren und in der Anwendung dieser informationstechnischen Innovationen in zahlreichen Sektoren und Lebensbereichen (Sejnowski 2018).

Für den „algorithmic turn" in der Governance von Plattformen sind vor allem Fortschritte in der automatisierten Objekt- und Mustererkennung in Bildern, Videos und Audioinhalten relevant sowie die automatisierte Klassifikation von Texteinheiten. Dabei lassen sich grob zwei unterschiedliche Vorgehensweisen unterscheiden: Matching und Klassifikation (Gorwa et al. 2020, S. 3–5).

Matching beschreibt Verfahren, die darauf zielen, neue Inhalte gegen eine bekannte Menge von Inhalten abzugleichen, etwa um das illegale Posten von urheberrechtlich geschützten Musikstücken oder von bereits identifizierten terroristischen Inhalten zu verhindern. Diese Verfahren basieren in der Regel auf der Umwandlung eines bekannten Inhalts in einen „Hash", d. h. eine Zeichenfolge, die den zugrunde liegenden Inhalt eindeutig identifizieren soll. Hashes sind nützlich, weil sie einfach zu berechnen sind und in der Regel deutlich weniger Datenvolumen benötigen als der zugrundeliegende Inhalt. Das macht die Speicherung und den Vergleich deutlich effizienter. Im Unterschied zu klassischen, etwa kryptografischen Hashes zielen neuere Hash-Funktionen für die Moderation von Inhalten darauf, auch den Vergleich von leicht geänderten Inhalten zu ermöglichen (Engstrom und Feamster 2017). Ansonsten könnten diese automatisierten Systeme leicht durch kleine Änderungen oder Störungen (wie das Ändern von wenigen Pixeln, Hinzufügen von Wasserzeichen, Beschneiden der Ränder) umgangen werden. Diese neueren Verfahren berechnen also keine exakten Übereinstimmungen, sondern eher Homologien, also Ähnlichkeiten zwischen zwei Eingaben (Datar et al. 2004). Besonders relevant ist das sogenannte „perceptual hashing" (p-hash; Niu und Jiao 2008). Beim perzeptuellen Hashing werden bestimmte wahrnehmbare Merkmale von Inhalten, wie z. B. Ecken in Bildern oder Frequenzen in Audiodateien, erkannt und als charakteristischer „Fingerabdruck" des Inhalts identifiziert. „[E]ach note in a song could be represented by the presence or absence of specific frequency values; the volume of a particular note ... by some amplitude at the frequency corresponding to that musical note" (Duarte et al. 2017, S. 14). Indem diese semantischen Merkmale anstelle der in kryptografischen Hashes verwendeten Bitfolgen erfasst werden, ist dieses Fingerprinting robuster gegenüber Änderungen, die für die menschliche Wahrnehmung des Inhalts irrelevant sind. Gleichzeitig ist es besser geeignet, neue Variationen und Interpretationen eines Inhalts zu erkennen.

Im Unterschied zu diesen Varianten des Matching zielt das zweite Verfahren, die *Klassifikation,* auf die Charakterisierung und Kategorisierung von bislang unbekannten Inhalten. Diese Verfahren basieren in der Regel auf neueren Formen maschinellen

Lernens, d. h. der automatischen Erkennung von statistischen Mustern aus idealerweise großen Mengen an Daten. Etwa mit Blick auf Hassrede versuchen Anbieter mithilfe dieser Systeme, gewaltvolle Sprache nicht mehr anhand von Blacklists (also einer endlichen Liste von idenfizierten Schlüsselwörtern) zu erkennen, sondern durch die Identifikation komplexerer Muster von Wort- und Satzzusammenhängen auf Basis von Natural Language Processing (NLP) (Schmidt und Wiegand 2017). Aktuelle NLP-Verfahren nutzen meist „supervised learning“, also überwachtes Lernen, bei dem algorithmische Modelle durch Menschen trainiert werden (vgl. zur folgenden Darstellung Duarte et al. 2017, S. 8–10). Zunächst wird ein Trainingsdatensatz aufgebaut, in dem Inhalte entlang der interessierenden Merkmale bereits klassifiziert sind, also z. B. Hassrede/konflikthaft/keine Merkmale. Dieser Datensatz wird dann genutzt, um anhand von Verfahren der Erkennung von Textähnlichkeiten wie N-grams und Word-Embedding ein abstraktes Modell zu entwickeln (Mikolov et al. 2013), das die interessierenden Merkmale mit Textausschnitten in Verbindung setzt. Auf dieser Basis wird das Modell dann „trainiert“, indem es nun iterativ neue Inhalte nach diesem Modell klassifiziert und dabei von Menschen überwacht wird, die die Entscheidungen des Systems als richtig oder falsch markieren. Dadurch wird das System immer weiter angepasst und auf eine möglichst niedrige Fehlerquote entlang der vordefinierten Kategorien und des vorhandenen Datensatzes optimiert.

Diese Matching- und Klassifikationsverfahren wurden in den vergangenen 10 Jahren stark weiterentwickelt und begründen die heutigen Verfahren der automatisierten Inhaltemoderation. Das Urheberrecht war historisch vermutlich der erste Bereich, in dem – vor allem aufgrund starker wirtschaftlicher Interessen – Technologien zum automatisierten Abgleich und zur Klassifizierung von Online-Inhalten eingesetzt wurden. Als Ergebnis der Filesharing-Kontroversen in den frühen 2000er-Jahren hat sich in diesem Bereich auch der „responsibility turn“ und die engere Ausgestaltung des Haftungsprivilegs sehr viel früher gezeigt (Katzenbach 2017). So hat YouTube schon 2006 begonnen, mit Systemen zur Erkennung von urheberrechtlich Inhalten zu experimentieren (Holland et al. 2016). Diese Bemühungen entwickelten sich im Laufe der Zeit zu dem Content-ID-System, das YouTube nun schon seit mehr als einem Jahrzehnt betreibt und kontinuierlich weiterentwickelt. Auch wenn YouTube die spezifische technologische Implementierung seines proprietären Systems geheim hält, lassen sich Grundzüge des Systems darstellen. Es ermöglicht Inhabern von Urheberrechten, ihr Material in eine Datenbank hochzuladen, das (a) mit bestehenden Inhalten auf YouTube abgeglichen wird und (b) einer Hash-Datenbank hinzugefügt wird, um neue Uploads dieser Inhalte zu erkennen. Im Zusammenhang mit dem Urheberrecht besteht das Ziel des Einsatzes automatischer Systeme nicht nur darin, identische Dateien zu finden, sondern auch verschiedene Instanzen und Aufführungen kultureller Werke zu identifizieren, die möglicherweise urheberrechtlich geschützt sind. Durch die aktuellen Fingerprinting-Verfahren sind diese Systeme nicht nur in der Lage, mehrfache Uploads eines Musikvideos zu finden, sondern auch von Aufnahmen von Live-Aufführungen dieses Songs.

Längst entwickeln, testen und betreiben Plattformen aber automatisierte Systeme auch, um Hassrede zu identifizieren, terroristische Inhalte zu blocken und Desinformationskampagnen zu kennzeichnen (vgl. Tab. 1, sowie Gillespie 2018; Duarte et al. 2017). Das Global Internet Forum to Counter Terrorism (GIFCT), eine gemeinsa-

Tab. 1 Übersicht über eingesetzte automatisierte Verfahren nach Plattformen. (Gorwa et al. 2020, S. 6)

	Terrorism	Violence	Toxic speech	Copyright	Child abuse	Sexual content	Spam & automated accounts
Facebook	Shared Industry Hash Database (SIHD), ISIS/Al-Qaeda classifier	Community standards classifiers	Community standards classifiers	Rights manager	PhotoDNA	Non-consensual intimate image classifier, nudity detextion	Immune system
Instagram	–	–	Comment filter	Rights manager	PhotoDNA	–	Comment filter, false account detection
YouTube	SIHD, Community Guidelines (CG) ML classifiers	CG ML Classifiers	CG ML Classifiers	Content ID	Content safety API, PhotoDNA	CG ML Classifiers	CG ML Classifiers
Twitter	SIHD	–	Quality filter	–	PhotoDNA	Sexual content interstitial	Proactive Tweet and account detection, quality filter
WhatsApp	–	–	–	–	PhotoDNA	–	Modified immune system

API „application programming interface“

me Initiative von Facebook, Google, Twitter und Microsoft zur Bekämpfung der Verbreitung terroristischer Inhalte im Internet, betreibt etwa eine gemeinsame, aber geheime Datenbank mit identifizierten terroristischen Bildern, Videos, Audios und Texten. Hier werden also hauptsächlich Verfahren des Matching eingesetzt. Es wurde in frühen Pressemitteilungen betont, dass „übereinstimmende Inhalte nicht automatisch entfernt werden“ (Facebook Newsroom 2016). Die Reaktionen der Plattformen auf Vorfälle wie die Schießerei im neuseeländischen Christchurch und auf Propaganda großer Terrororganisationen wie ISIS und Al-Qaida scheinen allerdings darauf hinzudeuten, dass Treffer der GIFCT-Datenbank inzwischen durchaus automatisch blockiert werden, ohne dass menschliche Moderatoren involviert sind.

7 Probleme und Herausforderungen im Einsatz algorithmischer Systeme in der Governance von Plattformen

Die algorithmische Wende in der Governance von Plattformen ist also vielschichtig. Fraglos vorhandene technische Fortschritte in der Identifikation und Klassifikation von Inhalten sind durch die Verbindung mit allgemeinen Technikutopien, dem Motiv des „technological fix“ und wachsendem politisch-regulativem Druck erfolgreich als vielversprechende Lösungen für komplexe Herausforderungen positioniert worden. Plattformanbieter, aber auch politische Institutionen, versprechen sich auf diese Weise Fortschritte in der Bekämpfung von Terrorismus und Kindesmissbrauch im Netz, von Misinformation und Hassrede. Der flächendeckende Einsatz automatisierter Systeme zur Adressierung dieser komplexen Herausforderungen birgt aber zahlreiche praktische und grundsätzliche Herausforderungen und Probleme.

Eine erste Gruppe von Problemen betrifft den Umstand, dass diese Systeme weit davon entfernt sind, perfekt zu funktionieren. Mit anderen Worten: In der Praxis ist *die fehlerhafte Klassifikation von Inhalten* kein seltener Fall. So werden Inhalte fälschlicherweise als problematisch oder illegal markiert und gegebenenfalls blockiert („false positives“); oder andersherum tatsächliche illegale Inhalte wiederum nicht erkannt („false negatives“). Plattformen berichten in ihren Transparenzberichten zwar regelmäßig hohe und steigende Trefferquoten für terroristische Inhalte und Hassrede. Im „Community Guidelines Enforcement Report“ berichtet Facebook etwa für das erste Quartal 2021, dass knapp 97 % der auf der Plattform identifizierten Hassrede proaktiv durch das Unternehmen erkannt wurde, fast ausschließlich durch automatisierte Verfahren. Zwei Jahre zuvor waren das nur 67 %. Das spricht für einen deutlich wachsenden Einsatz dieser Technologien, sagt aber wenig über deren Präzision aus. Wie viele „false positives“ sind hier mit enthalten (Overblocking), wie viele „false negatives“ fehlen in dieser Liste (Underblocking)?

Unabhängige empirische Forschung zu diesen Fragen gibt es wenig, vor allem weil es kaum reliable Daten gibt. Für den Urheberrechtsbereich weisen die wenigen vorhandenen Studien auf deutliches Overblocking hin. Auch wenn Content ID und andere Systeme aus technischer Sicht inzwischen sehr elaborierte Fingerabdrücke für Inhalte erstellen, bedeutet dies nicht zwangsläufig, dass sie bei der Bewertung tatsächlicher Urheberrechtsverletzungen sehr viel besser geworden sind. Denn das Urheberrecht erlaubt es Dritten durchaus, Auszüge aus geschützten Werken im Rah-

men der Schrankenregeln oder „Fair-use"-Klauseln zu verwenden. Empirische Studien zur automatisierten Durchsetzung von Urheberrechten kommen so vielleicht nicht überraschend zum Befund des Overblockings: Inhalte werden systematisch zu häufig ausgefiltert, obwohl sie aus urheberrechtlicher Sicht eigentlich veröffentlicht werden dürften (Bar-Ziv und Elkin-Koren 2018; Erickson und Kretschmer 2018; Urban et al. 2017). Burk und Cohen (2001) haben schon früh argumentiert, dass die zahlreichen *kontextuellen Faktoren,* die für die Bewertung von „Fair-use"-Standards und Schrankenregeln erforderlich sind, grundsätzlich nicht in automatisierte Systeme einprogrammiert werden können. Diese hohe Kontextabhängigkeit in der Einordnung und Bewertung von Inhalten ist ein Faktor, der nicht nur für urheberrechtliche Fragen gilt, sondern auch bei Hassrede und Misinformation in hohem Maße relevant ist. Ein identischer Satz kann je nach Umfeld als Hassrede verstanden werden (und gedacht sein!) oder nicht. So veranschaulichen Duarte et al. (2017, S. 19): „Often, context and minor semantic differences seperate hate speech from benign speech. For example, the term 'slant' is a slur often used to insult the appearance of people of Asian descent, but 'The Slants' is an Asian-American band whose members chose the name in part to undercut slurs about Asian-Americans that band members heard in childhood." Und der Wahrheitsgehalt einer Information hängt ohnehin vom Kontext ab. Automatisierte Verfahren können aber solche Kontextfaktoren nicht mitberücksichtigen, sondern operieren in der Regel allein auf Basis des fraglichen Inhalts.

Aber selbst, wenn man sich vorstellt, automatisierte Systeme würden perfekt entlang der Vorgaben und Kriterien verfahren und Inhalte klassifizieren, bleiben grundsätzliche Probleme und Herausforderungen bestehen, die nicht durch technische Optimierung gelöst werden können. Mit Blick auf allgemeine Probleme des Einsatzes von KI und Automatisierung wird bereits intensiv über systematische Diskriminierung und Benachteiligung von ohnehin marginalisierten Gruppen diskutiert. Zahlreiche Studien zur Automatisierung von Prozessen in Bereichen wie Kreditvergabe, Einstellungsverfahren und Gerichts- und Polizeiarbeit zeigen, dass algorithmische Systeme routinemäßig Menschen und Kollektive begünstigen, die bereits privilegiert sind, während marginalisierte Menschen systematisch diskriminiert und benachteiligt werden (Barocas und Selbst 2016; Berk et al. 2018; Noble 2018). Dies gilt auch für die Moderation von Social-Media-Inhalten: Klassifikationssysteme für Empfehlungen, Rankings oder Sperrungen auf Plattformen verfestigen Formen der Ungleichheit und begünstigen Mehrheitsgruppen. So werden beispielsweise Klassifikatoren für Hassrede, die darauf ausgelegt sind, Verstöße gegen die Richtlinien einer Plattform zu erkennen, Sprache, die von einer bestimmten sozialen Gruppe verwendet wird, unverhältnismäßig stark markieren, sodass die Äußerungen dieser Gruppe mit größerer Wahrscheinlichkeit entfernt werden (Binns 2018; Blodgett et al. 2016; Ekstrand et al. 2018). Bei diesen Fragen geht es allerdings zuletzt nicht einfach um „algorithmic fairness", die wiederum technisch optimiert werden könnte (Binns et al. 2017). Sondern dahinter liegen grundlegende soziale Ungleichheiten, die nicht erst durch Plattformen und ihre Algorithmen begründet wurden – gegenwärtig von diesen aber massiv verstärkt werden (Hoffmann 2019; Dencik et al. 2019).

Ein weiteres grundlegendes Problem der algorithmischen Wende in der Governance von Plattformen ist *mangelnde Transparenz* der Entscheidungskriterien und -verfahren. Während die notorisch intransparenten Moderationsregeln und -verfahren der Plattformen durch intensive empirische Forschung (Roberts 2019) und die dank massiven öffentlichen Drucks angestoßenen Aktivitäten aufseiten der Anbieter – Transparenzberichte, Veröffentlichung von Moderationsregeln, Aufsichtsmechanismen wie Facebooks Oversight Board (Suzor et al. 2019; Klonick 2020) – langsam besser beobachtbar werden, führen automatisierte Systeme nun eine erneute Erhöhung von Komplexität und Intransparenz ein. Die Dynamik von Sperrungen ist deutlich schwieriger von außen zu rekonstruieren, wenn Inhalte automatisiert als problematisch oder illegal markiert werden, vielleicht gar nicht erst veröffentlicht werden. Wie Llanso (2019) beschreibt, scheint es auch eine schleichende Ausweitung bestimmter algorithmischer Moderationssysteme zu geben, bei denen Unternehmen mit dem Hinzufügen neuer Funktionalitäten experimentieren – wie etwa die Ankündigung der GIFCT-Unternehmen, dass sie mit der gemeinsamen Nutzung von URL-Blockierlisten experimentieren – und zwar in der Regel ohne Aufsicht und Transparenz. Dabei geht es hier um fundamentale Fragen von Meinungs- und Informationsfreiheit: Welche Regeln und Verfahren entscheiden darüber, was öffentlich geäußert werden kann und wo die Grenzen liegen?

Vor diesem Hintergrund wird eine mittelfristige Herausforderung darin bestehen, diese grundlegenden politischen Fragen nicht in den vielleicht zukünftig scheinbar reibungslos operierenden KI-Systemen verschwinden zu lassen. Auch wenn wir derzeit recht intensiv über Fragen der Inhaltemoderation und der Governance durch Plattformen diskutieren, so muss das keineswegs immer so bleiben. Der gegenwärtige Institutionalisierungsprozess von Plattformen und ihrer Governance (Katzenbach 2021) kann sich durchaus mittelfristig auch stabilisieren und auf Basis von dann etablierten KI-Systemen der Inhalte-Klassifikation zu einer umfassenden *De-Politisierung* der grundlegenden Fragen von Meinungsfreiheit und Informationsvielfalt führen. Ist der „technological fix“ erstmal geglückt (d. h. als selbstverständlich etabliert, unabhängig von seinen tatsächlichen Effekten), verschwinden diese Fragen tendenziell aus dem Raum der Auseinandersetzung und werden zu unhinterfragten Infrastrukturen der öffentlichen Kommunikation. Mittelfristig wird die algorithmische Moderation und Regulierung immer nahtloser in unsere persönlichen und gesellschaftlichen Kommunikationsroutinen integriert sein. Vor diesem Hintergrund wird es eine Herausforderung für Beobachter, Forscherinnen und Aktivisten sein, einer solchen Normalisierung von automatisierten Entscheidungen über öffentliche Kommunikation entgegen zu wirken und Fragen der Inhaltemoderation auf der öffentlichen und politischen Agenda zu halten.

8 Fazit

Dieser Beitrag ist von der Beobachtung ausgegangen, dass automatisierte Systeme in Form von Algorithmen und KI-gestützten Verfahren zunehmend zur Moderation von Inhalten auf Plattform eingesetzt werden. Dabei werden sie regelmäßig sowohl von Plattformbetreibern wie auch von politischen Akteuren als zentrale Lösung für die

systematischen Probleme von Plattformen mit Blick auf Hassrede, Misinformation und andere schadvolle Inhalte positioniert. Der Beitrag hat gezeigt, dass dieser „algorithmic turn“ in der Regulierung und Moderation von Online-Inhalten nicht allein eine technische Entwicklung ist, sondern stark von diskursiven und politischen Entwicklungen gestützt und getrieben wird. Mit anderen Worten: Algorithmen und KI erscheinen nicht deshalb als adäquate Lösung, weil die technologische Entwicklung so weit fortgeschritten sei, sondern weil der öffentliche und politische Druck auf Plattform so gewachsen ist, dass sie eine Lösung präsentieren müssen – und gleichzeitig das bewährte Motiv des „technological fix“ für eine solche Positionierung eine gute diskursive Grundlage bietet.

Tatsächlich ist diese algorithmische Wende aber mit eigenen, teils ganz grundsätzlichen Problemen verbunden, wie der Beitrag gezeigt hat: weiterhin hohe technische Fehlerraten; grundsätzliche Grenzen in der notwendigen Berücksichtigung von Kontext in der Bewertung von Inhalten; Verstärkung der systematischen Benachteiligung von marginalisierten Gruppen; mangelnde Transparenz; sowie die De-Politisierung von gesellschaftlich konstitutiven Fragen wie nach der Abwägung von Meinungsfreiheit und Persönlichkeitsrechten. Zuckerbergs „AI will fix this“ ist also ganz sicher nicht die Antwort auf die Probleme von Plattformen und die damit verbundenen gesellschaftlichen Herausforderungen.

Tatsächlich scheint sich hier so etwas wie ein Plattformparadox entwickelt zu haben: Einerseits hat sich ein breiter gesellschaftlicher Konsens darüber entwickelt, dass die großen Plattformen mehr Verantwortung übernehmen müssen und ihre Macht zu begrenzen ist. Andererseits scheinen die vorhandenen Regulierungsversuche letztlich eher die Meinungsfreiheit zu begrenzen und den großen Plattformen im Gegenzug noch immer mehr Macht zuzuweisen. Allein die Zuweisung von wachsenden Haftungsansprüchen an Plattformen wie im deutschen NetzDG und der Europäischen Urheberrechtsrichtlinie inzentiviert die Plattformen, Inhalte vorsorglich zu sperren und so potenziell Meinungsfreiheiten für Nutzer zu reduzieren. Letztlich wächst damit sogar die Macht der großen Plattformen, da sie gedrängt werden, noch mehr Verfügungsgewalt über Inhalte auszuüben.

Ein Ausweg aus diesem Plattformparadox ist noch nicht etabliert. Zwei Richtungen für einen gemeinwohlorientierten und erfolgversprechenden gesellschaftlichen und regulatorischen Umgang mit Plattformen lassen sich aber schon erkennen: Es scheint, dass (1) der Anspruch an die Übernahme von Verantwortung für Inhalte sich sinnvoller über durchsetzungsstarke Rechenschafts- und Transparenzrichtlinien gegenüber Plattform umsetzen lässt als durch enge Haftungsregeln. Die im Kontext des DSA auf EU-Ebene diskutierten Maßnahmen schlagen hier mit verpflichtenden Audits und daran obligatorisch anknüpfende Reaktionen der Plattformbetreiber den richtigen Weg ein. Auf nationaler Ebene in Deutschland weisen auch der neue § 5a des NetzDG und § 19 (3) des neuen Urheberrechts-Diensteanbieter-Gesetzes in eine gute Richtung: Diese Klauseln verpflichten Plattformen, Forschern Zugriff auf interne Daten zur Verfügung zu stellen, um die Auswirkungen der Inhaltemoderation von Plattform systematisch und unabhängig untersuchen zu können.

In der kritischen Beobachtung dieser Entwicklungen müssen wir als Forscher uns aber auch fragen, ob wir (2) nicht schon mit dieser Rahmung der Herausforderungen zu sehr den Narrativen der großen Plattformen folgen. Es scheint selbstverständlich,

dass die unvorstellbaren Mengen an Inhalten nur (fast) vollautomatisiert zu handhaben sind. Tarleton Gillespie (2020, S. 4) erinnert uns aber ganz richtig: „Still, every time we try to ‘solve’ this problem [of content moderation at scale], we pass up the opportunity to ask a much more profound question. ... Perhaps, if moderation is so overwhelming at this scale, it should be understood as a limiting factor on the ‘growth at all costs’ mentality.“ Denn tatsächlich: Die rasante Skalierung ist eine der zentralen Erfolgsgeschichten der US-Tech-Szene. Nutzen und Erträge können in digital vernetzen Märkten (und besonders bei Plattformen) exponentiell wachsen, so heißt es, ohne dass die Kosten und Aufwände für das Unternehmen substanziell steigen. Ob wachsende Nutzerzahlen auf Facebook oder Millionen neue Fahrer für Uber – Serverfarmen, Algorithmen und Datenstrukturen nutzen das Wachstum zur Verbesserung der Dienste und erzeugen kaum neue Kosten (vgl. zu diesen Effekten u. a. Dogruel und Katzenbach 2018). Mit der immer stärker werdenden Erwartung an die Plattformen, Verantwortung für ihre Dienste und Angebote zu übernehmen, geht hier aber eine Veränderung einher. Genau betrachtet lässt es sich schwerlich vorstellen, wie Plattformen konsequent und gleichzeitig kontextsensibel wirklich Verantwortung für die (Vermittlung von) Kommunikation von Milliarden von Menschen übernehmen können. Hier könnten sich also die Skalierungsvorteile zumindest in einer Hinsicht ins Gegenteil verkehren – und damit könnte das scheinbar grenzenlose Wachstum der Plattformen auch für die Anbieter selbst zum Problem werden. Verantwortung und die kontextsensible Moderation von Inhalten skalieren nämlich überhaupt nicht gut. Die Regulierung der EU in den Entwürfen zum DSA setzen mit der asymmetrischen Regelsetzung und deutlichen strengeren Vorgaben an „sehr große Plattformen“ ebenfalls an dieser Dimension an.

Damit sind aber keineswegs alle Fragen beantwortet. Wir befinden uns derzeit in einer Phase tiefgreifender Veränderungsprozesse und damit verbunden einer politischen und gesellschaftlichen Suchbewegung, wie Plattformen zu regulieren und gemeinwohlorientiert zu gestalten sind. Und das ist auch nicht verwunderlich, denn Plattformen befinden sich derzeit selbst noch in einem Formationsprozess. Es ist keinesfalls klar, als welche Art von gesellschaftlicher Institution sie sich schließlich etablieren werden (Jarren 2019; Katzenbach 2021). Als neuer Typ von Vermittlungsinstitution lassen sie sich nicht einfach in bestehende Regulierungs- und Erwartungsstrukturen einfügen. Herauszuarbeiten, in welcher Form sich Plattformen gesellschaftlich gestalten und nachhaltig nutzen lassen, wird im Bereich von Medien und Kommunikation die zentrale Aufgabe der 2020er-Jahre sein.

Funding Open Access funding enabled and organized by Projekt DEAL.

Die diesem Artikel zugrundeliegende Forschung wurde mit Mitteln aus dem Forschungs- und Innovationsprogramm Horizont 2020 der Europäischen Union unter der Vereinbarungsnummer 870626 („ReCreating Europe“) und von der Deutschen Forschungsgemeinschaft (DFG) im Rahmen des multinationalen Programms „Open Research Area“ (ORA) unter der Förderungsnummer 440899634 („Shaping AI“) gefördert.

Declaration of Conflicting Interests Der Autor ist Mitglied des wissenschaftlichen Beirats „Actor & Behavior Policy" von Facebook/Meta. Er erhält dafür eine jährliche Aufwandsentschädigung von 6000 US$.

Literatur

Ames, Morgan G. 2018. Deconstructing the algorithmic sublime. *Big Data & Society* 5(1):1–4.

Bar-Ziv, Sharon, und Niva Elkin-Koren. 2018. Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown. *Connecticut Law Review* 50:339.

Barocas, Solon, und Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104(3):671.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns und Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50(1):3–44.

Binns, Reuben. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 81:149–159.

Binns, Reuben, Michael Veale, Max Van Kleek und Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, 405–415. Berlin: Springer.

Bloch-Wehba, Hannah. 2020. Automation in Moderation. *Cornell International Law Journal* 53:41–96.

Blodgett, Su Lin, Lisa Green und Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *EMNLP 2016: Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, USA, 1–5 November 2016.

Bory, Paolo. 2019. Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Convergence: The International Journal of Research into New Media Technologies* 25(4):627–642.

Brennen, J. Scott, Philip N. Howard und Rasmus Kleis Nielsen. 2018. An Industry-Led Debate: How UK Media Cover Artificial Intelligence. *Reuters Institute for the Study of Journalism*: 10.

Burk, Dan L., und Julie E. Cohen. 2001. Fair Use Infrastructure for Rights Management Systems. *Harvard Journal of Law & Technology* 15(1):41–83.

Campolo, Alexander, und Kate Crawford. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6:1.

Cave, Stephen, und Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1(2):74.

Citron, Danielle Keats, und Benjamin Wittes. 2017. The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity. *Fordham Law Review 86* (2): 401–423.

Datar, Mayur, Nicole Immorlica, Piotr Indyk und Vahab S. Mirrokni. 2004. Locality-Sensitive Hashing Scheme Based on p-Stable Distributions. In: *Proceedings of the twentieth annual symposium on computational geometry*, 2004:253–262. New York, NY: ACM.

Daub, Adrian. 2020. What Tech Calls Thinking: An Inquiry Into the Intellectual Bedrock of Silicon Valley. New York: Farrar Strauss & Giroux.

Degele, Nina. 2002. *Einführung in die Techniksoziologie*. München: Fink**.**

Dencik, Lina, Aren Hintz, Joanna Redden und Emiliane Treré. 2019. Exploring Data Justice: Conceptions, Applications and Directions. *Information, Communication & Society* 22(7):873–881.

DiMaggio, Paul. 1998. The new institutionalisms: avenues of collaboration. *Journal of Institutional and Theoretical Economics* 154(4):696–705.

Dogruel, Leyla, und Christian Katzenbach. 2018. Internet-Ökonomie: Grundlagen und Strategien und die Bedeutung von Plattformen beim Wirtschaften mit Medien- und Kommunikationsangeboten im Internet. In *Handbuch Online-Kommunikation,* Hrsg. Wolfgang Schweiger und Klaus Beck, 106–129. 2. Auflage. Wiesbaden: Springer.

Donges, Patrick. 2006. Medien als Institutionen und ihre Auswirkungen auf Organisationen. Perspektiven des soziologischen Neo-Institutionalismus für die Kommunikationswissenschaft. *Medien & Kommunikationswissenschaft* 54:563–578.
Duarte, Natasha, Emma Llanso und Anna Loup. 2017. Mixed Messages? The Limits of Automated Social Media Content Analysis. Washington, DC: *Center for Democracy & Technology*. https://perma.cc/NC9B-HYKX (Zugegriffen: 31. Mai. 2021).
Eisenegger, Mark. 2021. Dem digitalen Strukturwandel der Öffentlichkeit auf der Spur – Zur Einführung. In *Digitaler Strukturwandel der Öffentlichkeit. Historische Verortung, Modelle und Konsequenzen*, Hrsg. Mark Eisenegger, Marlis Prinzing, Patrik Ettinger und Roger Blum, 1–14. Wiesbaden: Springer VS.
Ekstrand, Michael D., Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan und Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*. New York, NY: ACM.
Elgan, Mike. 2017. Why fake news is a tech problem. With fake news wrecking everything, Silicon Valley is our last hope. *Computer World.* https://www.computerworld.com/article/3162020/why-fake-news-is-a-tech-problem.html (Zugegriffen: 31. Mai. 2021).
Engstrom, Evan, und Nick Feamster. 2017. *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools*. London: Engine.
Erickson, Kristofer, und Martin Kretschmer. 2018. This Video is Unavailable: Analyzing Copyright Takedown of User-Generated Content on YouTube. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 9(1).
Facebook Newsroom. 2016. Partnering to Help Curb Spread of Online Terrorist Content. https://perma.cc/V8DZ-AZZ7 (Zugegriffen: 31. Mai. 2021).
Fischer, Sarah, und Cornelius Puschmann. 2021. *Wie Deutschland über Algorithmen schreibt. Eine Analyse des Mediendiskurses über Algorithmen und Künstliche Intelligenz (2005–2020).* 1. Ausgabe. Gütersloh: Bertelsmann Stiftung.
Gasser, Urs, und Wolfgang Schulz. 2015. Governance of Online Intermediaries: Observations from a Series of National Case Studies. *Berkman Center Research Publication* 2015(5).
Gillespie, Tarleton. 2010. The Politics of ‘Platforms’. *New Media & Society* 12(3):347–364.
Gillespie, Tarleton. 2014. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, Hrsg. Tarleton Gillespie, Pablo Boczkowski und Kirtsen A. Foot, 167–193. Cambridge, MA: MIT Press.
Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* New Haven/ London: Yale University Press.
Gillespie, Tarleton. 2020. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2): 1–5.
Gorwa, Robert, Reuben Binns und Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* (January 2020).
Hall, Peter A., und Rosemary C. R. Taylor. 1996. Political Science and the Three New Institutionalisms. *Political Studies* 44(5):936–957.
Hasse, Raimund, und Georg Krücken. 2005. *Neo-Institutionalismus*. Bielefeld: transcript.
Haupt, Joachim. 2021. Facebook Futures: Mark Zuckerberg’s Discursive Construction of a Better World. *New Media & Society* 23(2):237–257.
Hoffmann, Anna Lauren. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22(7):900–915.
Holland, Adam, Chris Bavitz, Jeff Hermes, Andy Sellars, Ryan Budish, Michael Lambert und Nick Decoster. 2016. Intermediary Liability in the United States. *Berkman Centre for Internet & Society NOC Case Study Series*. https://perma.cc/2QAY-UTDY (Zugegriffen: 31. Mai 2021).
Jarren, Otfried. 2019. Fundamentale Institutionalisierung: Social Media als neue globale Kommunikationsinfrastruktur. *Publizistik* 64(2):163–79.
Katzenbach, Christian. 2012. Technologies as Institutions: Rethinking the Role of Technology in Media Governance Constellations. In *Trends in Communication Policy Research: New Theories, New Methods, New Subjects*, Hrsg. Natascha Just und Manuel Puppis, 117–138. Bristol, UK: Intellect Books.
Katzenbach, Christian. 2017. *Die Regeln digitaler Kommunikation. Governance zwischen Norm, Diskurs und Technik.* Wiesbaden: Springer VS.
Katzenbach, Christian. 2021. Die Öffentlichkeit der Plattformen: Wechselseitige (Re-)Institutionalisierung von Öffentlichkeiten und Plattformen. In *Digitaler Strukturwandel der Öffentlichkeit. Historische*

Verortung, Modelle und Konsequenzen, Hrsg. Mark Eisenegger, Marlis Prinzing, Patrik Ettinger und Roger Blum, 65–80. Wiesbaden: Springer VS.

Katzenbach, Christian. 2019. AI will fix this. In *Busted! The Truth About the 50 Most Common Internet Myths,* Hrsg. Matthias C. Kettemann und Stephan Dreyer, 194–195. Internet Governance Forum Berlin: Leibniz Institute for Media Research | Hans-Bredow-Institute. https://www.hiig.de/wp-content/uploads/2019/11/dgzmogf_KettemannDreyerInternetMyths2019-1.pdf (Zugegriffen: 31. Mai 2021).

Klonick, Kate. 2020. Creating Global Governance for Online Speech: The Development of Facebook's Oversight Board. *Yale Law Journal* 129(8):2418–2499.

Kuczerawy, Aleksandra, und Jef Ausloos. 2015. From Notice-and-Takedown to Notice-and-Delist: Implementing Google Spain. *Colorado Technology Law Journal* 14(2):219–258. CiTiP Working Paper 24.

Künzler, Matthias, Franziska Oehmer, Manuel Puppis und Christian Wassmer. 2013. *Medien als Institutionen und Organisationen: Institutionalistische Ansätze in der Publizistik- und Kommunikationswissenschaft*. Baden-Baden: Nomos.

Lischka, Juliane A. 2019. Strategic Communication as Discursive Institutional Work: A Critical Discourse Analysis of Mark Zuckerberg's Legitimacy Talk at the European Parliament. *International Journal of Strategic Communication* 13(3):197–213.

Llanso, Emma. 2019. Platforms Want Centralized Censorship. That Should Scare You. Opinion: Controlling the spread of insidious content online is extremely difficult—but combining efforts across platforms raises serious threats to free expression. *Wired.* https:// perma.cc/PWB3-NU3Q (Zugegriffen: 31. Mai 2021).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado und Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *27th Conference on Neural Information Processing Systems (NeurIPS)*. Lake Tahoe, NV.

Morozov, Evgeny. 2011. Don't be evil. *The New Republic*. http://www.newrepublic.com/article/books/magazine/91916/google-schmidt-obama-gates-technocrats. Zugegriffen: 31. Mai 2021.

Morozov, Evgeny. 2013. *To save everything, click here: The folly of technological solutionism*. New York: PublicAffairs.

Napoli, Philip M. 2013. *The Algorithm as Institution: Toward a Theoretical Framework for Automated Media Production and Consumption*. New York, NY, USA: Fordham University Schools of Business Research Paper.

Napoli, Philip M., und Robyn Caplan. 2017. Why Media Companies Insist They're Not Media Companies, Why They're Wrong, and Why It Matters. *First Monday* 22(5).

Niu, Xia-Mu, und Yu-Hua Jiao. 2008. An Overview of Perceptual Hashing. *Chinese Journal of Electronics* 36(7):1405-1411.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism.* 1 Ausgabe. New York: NYU Press.

Puppis, Manuel. 2010. Media Governance: A New Concept for the Analysis of Media Policy and Regulation. *Communication, Culture & Critique* 3(2):134–49.

Righetti, Nicola. 2021. Four Years of Fake News. A Quantitative Analysis of the Scientific Literature. *SocArXiv* 11 March.

Roberts, Sarah T. 2019. *Behind the Screen. Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

Russell, Frank Michael. 2019. The New Gatekeepers: An Institutional-level View of Silicon Valley and the Disruption of Journalism. *Journalism Studies* 20(5):631–648.

Sandhu, Swaran. 2018. Kommunikativer Institutionalismus und Accounts. In *Handbuch Sprache in den Public Relations*, Hrsg. Annika Schach und Cathrin Christoph, 21–36. Wiesbaden: Springer Fachmedien Wiesbaden.

Schmidt, Anna, und Michael Wiegand. 2017. *A Survey on Hate Speech Detection using Natural Language Processing*. Valencia, Spain: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 3 April 2017.

Schulz, Wolfgang. 2019. *Roles and Responsibilities of Information Intermediaries. Fighting Misinformation as a Test Case for a Human Rights-Respecting Governance of Social Media Platforms.* Hoover Institution, Stanford University, USA: Aegis series Paper 1904.

Schulz, Wolfgang. 2018. Regulating Intermediaries to Protect Privacy Online—the Case of the German NetzDG. *HIIG Discussion Paper Series, 2018–01*. https://www.hiig.de/wp-content/uploads/2018/07/SSRN-id3216572.pdf (Zugegriffen: 31. Mai. 2021).

Scott, W. Richard. 2008. *Institutions and Organizations: Ideas and Interests.* 3. Ausgabe. Los Angeles: SAGE.

Seetharaman, Deepa. 2016.Facebook Looks to Harness Artificial Intelligence to Weed Out Fake News. Company executives say the social network first needs a policy on how to responsibly apply such capabilities. *Wall Street Journal.* https://www.wsj.com/articles/facebook-could-develop-artificial-intelligence-to-weed-out-fake-news-1480608004?mod=pls_whats_news_us_business_f (Zugegriffen: 31. Mai 2021).

Sejnowski, Terrence J. 2018. *The Deep Learning Revolution*. Cambridge, MA, USA/ London, UK: MIT Press.

Suzor, Nicolas P., Sarah Myers West, Andrew Quodling und Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13(2019),1526–1543.

Síthigh, DaithíMac. 2020. The road to responsibilities: new attitudes towards Internet intermediaries*. *Information & Communications Technology Law* 29(1):1–21.

Tworek, Heidi J. S. 2021. Fighting Hate with Speech Law: Media and German Visions of Democracy. *The Journal of Holocaust Research* 35(2):106–122.

Urban, Jennifer M., Joe Karaganis und Brianna Schofield. 2017. *Notice and Takedown in Everyday Practice*. *Version 2*. Berkeley, CA, USA: UC Berkeley Public Law.

Vaidhyanathan, Siva. 2012. *The Googlization of Everything (And Why We Should Worry)*. 1. Ausgabe. Berkeley, CA, USA: University of California Press.

Volti, Rudi. 2014. *Society and Technological Change*. 1995. 7. Ausgabe. New York, NY, USA: Worth Publishers.

Christian Katzenbach 1979, Prof. Dr., Professor für Kommunikations- und Medienwissenschaft mit Schwerpunkt Medien-Governance und Plattformökonomie. Zentrum für Medien-, Kommunikations- und Informationsforschung (ZeMKI), Universität Bremen. Forschungsgebiete: Regulierung und Governance von Plattformen, Sozialen Medien und KI; Diskurse über neue Technologien und Medien; Techniksoziologie; Algorithmen und Automatisierung. Veröffentlichungen: Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. Science, Technology, & Human Values, 2021 (mit J. Bareis); Future imaginaries in the making and governing of digital technology. Special Issue. New Media & Society 23, 2021 (als Hrsg. mit A. Mager).