API

OPEN ACCESS

Publiziert im Juni 2023
DOI: 10.15460/apimagazin.2023.4.2.147

Zeitschriftenartikel

Begutachtet

Begutachtet:

Prof. Dr. Ulrike Verch
HAW Hamburg
Deutschland

Erhalten: *10. Mai 2023*
Akzeptiert: *25. Mai 2023*
Publiziert: *29. Juni 2023*

Empfohlene Zitierung:

HERRMANN, Bettina, 2022: Kann künstliche Intelligenz vor Hatespeech schützen? KI-gestützte Content Moderation auf Social-Media. In: *API Magazin* **4**(2) [Online] Verfügbar unter: DOI 10.15460/apimagazin.2023.4.2.147

Kann künstliche Intelligenz vor Hatespeech schützen?

KI-gestützte Content Moderation auf Social-Media

Bettina Herrmann[1]*

[1] Hochschule für Angewandte Wissenschaften Hamburg, Deutschland
Studentin im 2. Semester des Studiengangs Digitale Transformation der Informations- und Medienwirtschaft

* Korrespondenz: redaktion-api@haw-hamburg.de

Zusammenfassung

Content Moderation auf Social-Media-Plattformen wird aufgrund der großen Menge an Daten meist mit der Unterstützung von künstlicher Intelligenz durchgeführt. Das ist nötig, um schädliche Inhalte wie Hatespeech auf diesen Plattformen zu minimieren. Der Beitrag beschäftigt sich in Form eines Literaturreviews mit den Problemen, die der Einsatz von KI bei Content Moderation in diesem Feld mit sich bringen kann.

Schlagwörter: Content Moderation, Künstliche Intelligenz, Social Media, Hatespeech

Can artificial intelligence protect against hatespeech?

AI-supported content moderation on social-media

Abstract

Content Moderation is a vital part of social media platforms. It is a task usually supported by the use of artificial intelligence due to the size of most platforms. To minimize harmful content such as hate speech, a platform has to implement content moderation. This text reviews the literature on this subject and compiles problems that might arise when employing AI for content moderation.

Keywords: Content Moderation, Artificial Intelligence, Social Media, Hatespeech

DOI: 10.15460/apimagazin.2023.4.2.147

1 Einleitung

Im späten Oktober 2022 hat Elon Musk, unter anderem Chief Executive Officer von SpaceX und Tesla, den Kurznachrichtendienst Twitter aufgekauft. Bereits im Vorfeld fragten sich viele Nutzer*innen, welche Veränderungen das mit sich bringen würde. Die ersten Effekte zeigen sich bereits: Hatespeech sowie Falschinformationen haben binnen weniger Tage auf der Plattform zugenommen. Konzerne wie General Motors haben angekündigt, vorerst keine Werbung mehr auf Twitter schalten zu wollen.[1] Begründet liegt diese Reaktion vor allem in der Unsicherheit darüber, wie die neue Strategie von Twitter im Bereich Content Moderation aussehen wird. Während Musk zwar betonte, dass die Plattform nicht von Hassrede überflutet werden dürfe, ist er gleichzeitig ein großer Verfechter einer grenzenlosen Meinungsfreiheit und Gegner von Einschränkungen der Äußerungsfreiheit und wird von bekannten Verbreiter*innen von Falschinformationen wie dem ehemaligen US-Präsidenten Donald Trump bestärkt. Musk möchte für Twitter einen „Content Moderation Council" einrichten, der aus mehreren Mitgliedern bestehen und Regeln zur Regulierung des Contents bestimmen soll.[2] Die unmittelbaren Auswirkungen, welche die Twitter-Übernahme durch Musk hatte, zeigen die Signifikanz von Content Moderation. Teilweise sogar bezeichnet als wichtigster Service, den Social Media liefert (vgl. Barrett 2020, S. 3), soll Content Moderation sicherstellen, dass soziale Plattformen ihren Nutzer*innen Sicherheit bieten. Content Moderation beeinflusst durch ein solches Kuratieren allerdings auch die „wahrgenommene Realität" (Djeffal 2022, S. 179) der Nutzer*innen. Diese Arbeit beschäftigt sich mit diesem Thema unter der Leitfrage „Welche Probleme können bei dem Einsatz von KI-gestützter Content Moderation im Umgang mit Hatespeech auf Social-Media-Plattformen auftreten?" und nutzt dafür die Methode des traditionell-narrativen Literatur-Reviews (vgl. Efron & Ravid 2019, S. 21).

Hatespeech oder auch Hassrede ist auf Social Media allgegenwärtig und in der Regel gegen marginalisierte Gruppen gerichtet. Die gängigen sozialen Plattformen verbieten solche Inhalte, da Hassrede nicht nur das Wohlbefinden von Nutzer*innen beeinträchtigen, sondern auch Gewalttaten auslösen kann (vgl. Davidson et al. 2017, S. 512). In extremer Form geschah das in Myanmar, wo im Jahr 2017 mehr als 10.000 Rohingya, Mitglieder einer ethnischen und größtenteils muslimischen Minderheit, getötet wurden und Hunderttausende flüchten mussten. Hier wurde Facebook von den Vereinten Nationen eine Mitverantwortung zugeschrieben, da das Unternehmen die Hassrede gegen Rohingya auf der Plattform lange ignoriert hatte (vgl. Barrett 2020, S. 20). Die Minimierung von schädlichen Inhalten durch KI-gestützte Content Moderation zieht spezielle Probleme nach sich, die in dieser Arbeit betrachtet werden sollen. Zunächst wird unter Punkt 2 der Begriff der Content Moderation im Allgemeinen definiert, sowie auf die Besonderheiten beim Einsatz von künstlicher Intelli-

1 Siehe https://www.tagesschau.de/wirtschaft/unternehmen/twitter-veraenderung-101.html [Online, Zugriff am 10.11.2022].
2 Siehe https://www.spiegel.de/netzwelt/was-hat-elon-musk-mit-twitter-or-ja-so-etwas-in-der-art-ergibt-sinn-a-5a322dbd-982e-436d-93fd-209ea86f968b [Online, Zugriff am 10.11.2022].

 DOI: 10.15460/apimagazin.2023.4.2.147

genz eingegangen. Kapitel 3 gibt dann einen Überblick über allgemeine Probleme bei KI-gestützter Content Moderation. Im letzten Kapitel wird ein Fazit gezogen.

2 Content Moderation

2.1 Definition

Grimmelmann (2015, S. 42) definiert Content Moderation als „the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." Barrett (2020, S. 3) betont deren Wichtigkeit: „content moderation is one of the fundamental services social media offers – perhaps the most fundamental". Content Moderation ist Teil der Führungsstrategie einer Plattform. Es gibt ein Regelwerk, welches von Nutzer*innen der Plattform befolgt und dessen Einhaltung durch (meist speziell ausgebildete) Moderator*innen systematisch überprüft wird. Dadurch soll sowohl ein gewisser Qualitätsstandard der Inhalte gewährleistet, als auch die Menge an schädlichen Inhalten minimiert werden. Zudem kann Content gelöscht oder es können Nutzer*innen für eine gewisse Zeit gesperrt, beziehungsweise vollständig von der Plattform entfernt werden (vgl. Petricca 2020, S. 308).

Die Herangehensweise kann hierbei reaktiv oder proaktiv sein. In reaktiven Systemen werden Inhalte nur überprüft, nachdem sie, in der Regel von Nutzer*innen, gemeldet wurden. Dieser Vorgang wird Flagging genannt und gibt Nutzer*innen die Möglichkeit, selbst gestaltend am Moderationsprozess teilzunehmen (vgl. Gongane et al. 2022, S. 32). Proaktive Systeme kontrollieren Content selbstständig, entweder durch menschliche Moderator*innen oder automatisierte Filter (vgl. Llansó 2020, S. 1). Während es einst ausreichend war, Foren und ähnliche Online-Plattformen durch eine vergleichsweise geringe Anzahl an freiwilligen Nutzer*innen, die dann meist den Titel Administrator*in oder Moderator*in erhielten, betreuen zu lassen, ist mittlerweile die Menge an Content so groß, dass auf automatisierte Filter zurückgegriffen wird (vgl. Gorwa et al. 2020, S. 2). Darauf wird im Unterkapitel 2.2 näher eingegangen. Zusätzlich stellen Plattformen auch menschliche Moderator*innen an. In den frühen Tagen von Facebook reichte ein kleines In-House-Team aus. Heute wird diese Aufgabe outgesourct. Für Facebook und Instagram arbeiteten im Jahr 2020 rund 15.000 Moderator*innen, während bei Twitter mit 1.500 vergleichsweise wenige für diesen Zweck angestellt waren (vgl. Barrett 2020, S. 4 f). Diese Arbeitnehmer*innen führen ihren Job oft unter prekären Bedingungen und mit schwerwiegenden psychischen Folgen aus, wie etwa posttraumatischen Belastungsstörungen (vgl. Pinchevski 2022, S. 2). Sie konsumieren Massen an gewaltvollen und traumatisierenden Inhalten, damit die Social-Media-Nutzenden diese nicht sehen müssen (vgl. ebd., S. 5). Content Moderation wird von Plattformbetreiber*innen durchgeführt, um eine qualitativ hochwertige Plattform herzustellen und das Wohlbefinden ihrer Community zu sichern (vgl. Jiang et al. 2020, S. 13669). Die Umgebung soll für Nutzer*innen ansprechend sein, damit sie mehr Zeit auf der Plattform verbringen. Das ist wiederum attraktiv für Werbetreibende. Schlussendlich geht es also um

DOI: 10.15460/apimagazin.2023.4.2.147

einen monetären Nutzen (vgl. Jiménez Durán 2022, S. 3). Die Plattformen produzieren damit ein bestimmtes Image. Je nachdem, welche Interessen die Moderationsregeln verfolgen und wie streng sie sind, wird die Plattform attraktiv für eine bestimmte Zielgruppe. So stellte Kor-Sins (2021, S. 14f.) beispielsweise fest, dass Twitters Branding als Ort für politische Diskurse in Verbindung mit algorithmischer Moderation die Plattform wenig attraktiv für Anhänger der Alt-Right-Bewegung machte, während Reddits neutralere und weniger strenge Regeln diese Nutzer*innen eher anzogen.

2.2 KI-gestützte Content Moderation

Nachdem Content Moderation im Allgemeinen definiert wurde, soll nun auf KI-gestützte Content Moderation im Speziellen eingegangen werden. Gorwa et al. (2020, S. 3) definieren diese so: „algorithmic commercial content moderation [are] systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown)."

Wenn Plattformen wachsen, immer mehr Nutzer*innen gewinnen und somit die Menge an Content zu groß wird, um von menschlichen Moderator*innen kontrolliert zu werden, werden in der Regel automatisierte Systeme, beziehungsweise künstliche Intelligenz (KI) eingesetzt (vgl. ebd., S. 3). Neben der Masse an Content sind auch die negativen psychischen Auswirkungen des Moderationsjobs ein Grund für den Einsatz künstlicher Intelligenz (vgl. Petricca 2020, S. 310). Außerdem ist es kostengünstiger, automatisierte Filter zu nutzen, statt signifikant mehr Moderator*innen anzustellen (vgl. Siapera & Viejo-Otero 2021, S. 127). Gongane et al. (2022) klassifizieren manuelle, semi-automatisierte und automatisierte Content Moderation. Während manuelle Content Moderation komplett auf menschlichen Moderator*innen (wie bereits in 2.1. beschrieben) und Flagging von Posts durch Plattform-Nutzer*innen beruht, wird bei semi-automatischer Moderation das Flagging von Algorithmen übernommen. Der Content wird danach noch einmal von Menschen statt einer KI überprüft. Anders ist das bei vollständig automatisierter Moderation, wo der komplette Prozess von Algorithmen übernommen wird (vgl. Gongane et al. 2022, S. 32f.).

Im Kontext von Content Moderation werden verschiedene automatisierte Prozesse als KI bezeichnet, selbst wenn sie es streng genommen nicht sind (vgl. Llansó et al. 2020, S. 3). Beispielsweise sind Keyword-Filter, die Content mit bereits bekanntem, gesperrten Content auf einer Blacklist abgleichen, zwar ein automatisierter Prozess, aber keine KI. Sie werden teilweise trotzdem so bezeichnet (vgl. Gillespie 2020, S. 3). Oft wird KI- gestützte Content Moderation auch Filter, beziehungsweise automatischer oder automatisierter Filter, genannt (vgl. Llansó 2020, S. 2). Im Folgenden werden einige wichtige Begriffe aus dem Bereich KI und Content Moderation genannt und erklärt.

DOI: 10.15460/apimagazin.2023.4.2.147

Gorwa et al. (2020, S. 4f.) unterscheiden automatisierte Content Moderation in Systeme, die Klassifizieren (Classification) und solche, die Abgleichen (Matching). Matching-Systeme führen einen Prozess aus, der als Hashing bezeichnet wird. Dabei wird ein beispielhafter Inhalt in einen „Hash" umgewandelt – d.h. es wird ein einzigartiger Wert zugeordnet, der dazu dient, den Inhalt sicher erkennen zu können; sozusagen ein Fingerabdruck des Inhalts. Hashes sind in der Regel kleinere Datenmengen als der ursprüngliche Inhalt, weshalb sie leichter zu verarbeiten sind. Es gibt einen Datensatz an Hashes, mit dem neue Inhalte abgeglichen und gegebenenfalls gesperrt werden können. Diese kryptographischen Hashfunktionen haben den Nachteil, dass die Content Moderation durch eine leichte Abänderung des Inhalts leicht umgangen werden kann. Deshalb werden oft andere Hashmatching-Techniken angewendet, die eher nach Gemeinsamkeiten zwischen Inhalten suchen, anstatt lediglich passgenaue Kopien zu identifizieren. Hashmatching ist ein automatisierter Prozess, ohne sich aber maschinelles Lernen zunutze zu machen.

Des Weiteren gibt es Classification-Systeme. Solche Systeme ordnen neue Inhalte in eine Kategorie ein. Im Gegensatz zu Matching-Systemen, die nur Inhalte verarbeiten können, die bereits ein Gegenstück im Datensatz vorweisen, können Classification-Systeme auch unbekannte Inhalte verarbeiten. Die Kategorien, in die Inhalte dabei eingeordnet werden, beinhalten schon viele Beispiele, anhand derer das System Eigenschaften oder Muster erkennt. Viele angewendete Systeme gehören zur zweiten Kategorie, also zu den Classification- Systemen, und beruhen auf Algorithmen des maschinellen Lernens, oder auch Machine Learning (vgl. Dias Oliva et al. 2021, S. 701). Diese sind in der Lage, auf Basis von Erfahrung zu lernen und sich selbst zu verbessern (vgl. Gongane et al. 2022, S. 19). Dabei gibt es „supervised learning", also beaufsichtigtes Lernen, und „unsupervised learning", also unbeaufsichtigtes Lernen. Bei der beaufsichtigten Form werden die Trainingsdaten für den Algorithmus annotiert; das heißt, es werden Kennzeichnungen wie zum Beispiel „Hatespeech" oder „keine Hatespeech" vergeben. Diese annotierten Datensätze sind besonders geeignet, um Hatespeech zu erkennen (vgl. Llansó et al. 2020, S. 4). Die Erstellung ist jedoch aufwändig und zeitintensiv, weshalb solche Trainingsdaten eher rar sind (vgl. Lee & Li 2020, S. 3). Beim unbeaufsichtigten Lernen wird ein Datensatz verwendet, der nicht annotiert ist. Das System erkennt hier Muster innerhalb der Daten selbst (vgl. Llansó et al. 2020, S. 4).

Natural Language Processing, oder auch Computerlinguistik, spielt eine wichtige Rolle in der automatisierten Content Moderation (vgl. ebd., S. 6). So werden Rechenmodelle bezeichnet, die menschliche Sprache verstehen sollen und sich statistischer Berechnungen sowie maschinellen Lernens bedienen (vgl. Otter et al. 2021, S. 604f.). Diese Technik verzeichnete in den letzten Jahren einen immensen Entwicklungsfortschritt (vgl. ebd., S. 618) und ist vor allem für die Erkennung von Hate Speech und Desinformation von Bedeutung (vgl. Gongane et al. 2022, S. 13).

DOI: 10.15460/apimagazin.2023.4.2.147

Weiterhin werden künstliche neuronale Netze (Artificial Neural Networks) genutzt (vgl. ebd., S. 12). Diese mehrstufigen Netze, die den neuronalen Verbindungen im menschlichen Gehirn nachempfunden sind, spielen eine wichtige Rolle in der Datenauswertung (vgl. Llansó et al. 2020, S. 4).

Zuletzt soll hier Deep Learning Erwähnung finden. Deep Learning ist ein Teilbereich des maschinellen Lernens, der sich künstliche neuronale Netze zunutze macht. Auch dieses Modell lernt eigenständig, tut dies im Gegensatz zum Machine Learning aber ohne menschliches Eingreifen (vgl. Gongane et al. 2022, S. 20).

Mit dem zunehmenden Wechsel zu KI-gestützter Content Moderation verabschieden sich Plattformbetreiber*innen von einem ex post Ansatz, also der nachträglichen Beurteilung von Inhalten, und bewegen sich zu einem ex ante Ansatz, indem Inhalte im Moment ihrer Veröffentlichung bewertet werden (vgl. Cobbe 2021, S. 741).

Der Einsatz von Automatisation macht Content Moderation nicht nur effizienter, sondern bestimmt auch die Ordnung, nach der eine Plattform funktioniert und geführt wird (vgl. Wright 2022, S. 9f.).

3 Probleme KI-gestützter Content Moderation

Wie bereits erwähnt, wird KI-gestützte Content Moderation eingesetzt, um Probleme wie die Überprüfung einer durch Menschen nicht zu bewältigenden Masse an Content oder die psychischen Folgen dieser Tätigkeit zu umgehen. Obwohl künstliche Intelligenz oft als „Retterin in der Not“ und Lösung all dieser Probleme gesehen wird (vgl. Gorwa et al. 2020, S. 2), bringt sie wiederum neue Probleme mit sich, die hier betrachtet werden sollen. Die Punkte in diesem Unterkapitel treten zwar auch im Allgemeinen mit KI-gestützter Content Moderation auf, sind aber insbesondere im Umgang mit Hatespeech von Relevanz.

3.1 Transparenz

Fehlende Transparenz von KI-gestützter Content Moderation ist ein Kernpunkt, der in der Literatur bemängelt wird. Da die gängigen sozialen Plattformen alle von gewinnorientierten und privatwirtschaftlichen Unternehmen betrieben werden, ist es für Forschende ein Anliegen, herauszufinden, welche Interessen durch kommerzielle Content Moderation verfolgt werden. Kommen Unternehmen lediglich ihrer sozialen Verantwortung nach oder gibt es andere Motive? Das ist nur schwer herauszufinden. Denn Content Moderation ist ein interner Prozess (vgl. Petricca 2020, S. 311) und Details über die genaue Funktionsweise werden beispielsweise bei Facebook durch Geheimhaltungsverträge vor der Öffentlichkeit verschlossen gehalten (vgl. Dachwitz & Reuter 2019, S. 60). Marshall (2021, S. 7) bezeichnet Facebooks Operationsweise als „double standards“: gegenüber den Nutzer*innen wird zwar

DOI: 10.15460/apimagazin.2023.4.2.147

kommuniziert, was auf der Plattform verboten ist, nämlich beispielsweise Hatespeech, aber wie die KI oder Moderator*innen zu dem Schluss kommen, was als Hatespeech eingeordnet wird, bleibt unklar. Das ist vor allem ein Problem, weil besonders ebensolche großen sozialen Plattformen signifikante Akteure des öffentlichen Raums sind. Durch Entscheidungen, welche Inhalte angezeigt werden und welche nicht, formen sie die private und öffentliche Kommunikation der Gesellschaft, die digital auf ebendiesen Plattformen stattfindet. Teilweise sind diese Plattformen das Hauptkommunikationsmittel für Nutzer*innen (vgl. Cobbe 2021, S. 740); besonders hier haben Moderationsentscheidungen potentiell großen Einfluss auf die wahrgenommene Realität und das Verhalten der Nutzer*innen (vgl. Dias Oliva et al. 2021, S. 702).

Auch für Personen, die mit Social Media ihren Lebensunterhalt verdienen, ist fehlende Transparenz ein Problem. YouTuber*innen verlieren die Orientierung und erleiden finanzielle Einbußen, wenn ihre Videos als Bestrafung für einen Verstoß gegen Plattformregeln entmonetarisiert werden. Daher entwickeln und teilen sie Techniken, wie die algorithmische Content Moderation umgangen und Strafen vermieden werden können (vgl. Ma und Kou 2021, S. 14). Ein Grund für die Geheimhaltung der Funktionsweise von Content Moderation ist unter anderem, dass Plattformen vermeiden möchten, sich einer umfassenden Rechenschaftspflicht unterstellen zu müssen, vor allem vor Gericht. Das würde den Moderationsprozess erheblich verlangsamen. Außerdem kann eine komplette Offenlegung der Moderationsregeln es einfacher machen, diese böswillig zu umgehen (vgl. Petricca 2020, S. 314). Mit dem zunehmenden Einsatz von künstlicher Intelligenz werden diese Entscheidungen, deren Regelgrundlage nicht komplett offenliegt, abermals undurchsichtiger (vgl. Gorwa et al. 2020, S. 11). Das leitet über zum nächsten Punkt.

3.2 Zensur und Redefreiheit

Intransparenz wird vor allem bemängelt, weil Content Moderation in Teilen als Beschneiden der Meinungsfreiheit oder sogar als Zensur eingestuft wird, was wiederum das öffentliche Interesse an der Offenlegung der Funktionsweise rechtfertigt (vgl. Petricca 2020, S. 312). Petricca (2020, S. 313) betont, dass Content Moderation keine Zensur sei. Denn werde eine soziale Plattform genutzt, stimme man auch den Nutzungsbedingungen zu – und willige somit ein, sich auch an die Regeln zu halten. Llansó (2020, S. 3f.) hingegen bezeichnet Content Moderation als „Prior restraint“, also eine Zensur, die vor dem Sprechen oder der Veröffentlichung von Aussagen ausgeübt werde; man müsse sich also eine Art Erlaubnis einer Autorität einholen. Demnach würden automatisierte Filter jeden Content als potenzielle Regelverletzung behandeln, während im echten Leben die Wahrscheinlichkeit, dass eine getätigte Aussage zur Anzeige gebracht und bestraft wird, ungleich kleiner sei. Dieser Unterschied im Grad der Überwachung forme das Erleben des Rechts auf Meinungsfreiheit.

DOI: 10.15460/apimagazin.2023.4.2.147

Bei Facebook träfen am Ende immer Menschen die Entscheidung über entfernte Inhalte, denn „[n]iemand soll den Eindruck bekommen, Maschinen würden über das hohe Gut der Meinungsfreiheit entscheiden“ (Dachwitz & Reuter 2019, S. 60). Dies ist nur ein kleiner Einblick in den Diskurs zur Beachtung von Meinungsfreiheit in diesem Bereich. Grundrechte müssen auch bei der Durchführung von Content Moderation geachtet werden. Verfahren zu deren Schutz sind wünschenswert (vgl. Löber 2022, S. 299).

3.3 Rechenschaftspflicht

Kurz soll hier auch auf den Themenkomplex Gesetze, Regulierungen und Rechenschaftspflicht eingegangen werden. Die bereits besprochenen Punkte Transparenz und Zensur erbitten die Frage, welche Rechenschaftspflicht für soziale Plattformen besteht. Viele Nationen haben eigene Gesetze, die den Schutz der Privatsphäre, die Bekämpfung von Hatespeech und Terrorismus und ähnlichem regeln. In Deutschland ist das 2017 in Kraft getretene NetzDG (Netzwerkdurchsetzungsgesetz, in Kraft getreten am 01.10.2017)[3] zu nennen, dass Plattformbetreiber*innen mit teils sehr hohen Geldstrafen droht, falls solche Inhalte nicht fristgerecht entfernt werden.

Globale, einheitliche Regeln gibt es allerdings nicht (vgl. Petricca 2020, S. 322f.). Seit dem 16.11.2022 gilt allerdings der Digital Services Act (DSA), der zumindest EU-weit eine einheitliche Regelung bietet. Unter Punkt 3.2. wird noch einmal genauer auf die uneinheitliche Gesetzeslage in Bezug auf Hatespeech eingegangen. Auch wenn sich soziale Plattformen an Gesetze halten müssen, üben sie durch algorithmische Content Moderation potenziell eine große Macht aus. Cobbe (2021, S. 761) bezeichnet die eingesetzten Algorithmen als „censorship systems“, die soziale Interaktionen und die Diskussionskultur von Nutzer*innen verändern und anhand von unternehmerischen Werten die Grenzen des Sagbaren bestimmen. Systemen, die künstliche Intelligenz nutzen, wird deshalb eine große, aber oft übersehene Macht zugesprochen (vgl. Löber 2022, S. 293). Fraglich ist, wer diese Machtausübung kontrolliert und welche Instanz diese Kontrolle übernehmen kann bzw. sollte. Djeffal (2022, S. 191) nennt „die Regulierung von Inhalten […] eine Operation am offenen Herzen der Demokratie“ und fordert, dass diese Aufgabe weder alleinig privaten Unternehmen noch staatlichen Instanzen zufallen, sondern unter Mitwirkung die Zivilgesellschaft erfüllt werden sollte. Ein solches Vorgehen nennt sich „aktivierende Regulierung“ und soll Machtmissbrauch verhindern.

3.4 Zuverlässigkeit

Wie zuverlässig moderieren KI-gestützte Systeme Inhalte? Diese Frage ist aufgrund mangelnder Transparenz schwer zu beantworten. Forscher*innen haben in der Regel keinen Zugriff auf die Trainingsdaten, die große soziale Plattformen nutzen (vgl. Löber 2022, S. 293). Einige Plattformen veröffentlichen aber Zahlen zu der von ihnen

3 Siehe https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html [Online, Zugriff am 16.06.2023].

DOI: 10.15460/apimagazin.2023.4.2.147

durchgeführten Content Moderation. Im Folgenden werden drei Beispiele genannt. Facebook veröffentlicht regelmäßig einen Community Standards Enforcement Report.[4] Dort findet man Information zur Umsetzung von Content Moderation in verschiedenen Bereichen der Regelverletzung. Für Hatespeech geht daraus hervor, dass bei einem absteigenden Trend im dritten Quartal 2022 0,02 % der von Nutzer*innen aufgerufenen Posts Hassrede enthielten. Von allen Inhalten, die Hassrede enthielten, wurden 90,2 % von Facebooks automatisierten Technologien proaktiv erkannt, bevor sie von Nutzer*innen gesehen wurden. Das ist eine hohe Erfolgsrate.

Twitter veröffentlichte Daten für das Jahr 2021, aus denen aber nicht hervorgeht, wie viele schädliche Inhalte proaktiv erkannt und wie viele von Nutzer*innen gemeldet wurden.[5] Im Jahr 2019 wurden nur 50 % der Tweets, die gegen Regeln verstießen, proaktiv von einem automatisierten System erkannt (vgl. Barrett 2020, S. 11).

Google veröffentlicht ebenfalls einen Bericht zur Durchsetzung der Community-Richtlinien.[6] Dort wird gleich zu Beginn darauf hingewiesen, dass durch die Corona-Pandemie weniger Personal anwesend ist und somit verstärkt automatisierte Filter genutzt werden, was vermehrte Fehltreffer zur Folge haben kann. Die Daten zeigen, dass der Großteil der entfernten Videos zwischen Juli und September 2022 algorithmisch erkannt und nicht von Nutzer*innen geflaggt wurden, nämlich 94,5 %. Von den entfernten Inhalten wurden 2,8 % als Hatespeech klassifiziert.

4 Probleme KI-gestützter Content Moderation in Bezug auf Hatespeech

In diesem Kapitel soll auf Probleme eingegangen werden, die besonders im Zusammenhang mit der Moderation von Hatespeech auf sozialen Plattformen relevant sind. Zunächst wird der Begriff Hatespeech definiert.

Eine einheitliche Definition gibt es nicht, sie kann je nach Region, Land oder auch Kontext unterschiedlich ausfallen (vgl. Gongane et al. 2022, S. 5). Was Hassrede ist und was nicht, wird gesellschaftlich ausgehandelt, wobei nicht zwingend ein Konsens erreicht wird (vgl. Gillespie 2020, S. 3). Davidson et al. (2017, S. 512) definieren Hatespeech als „language that is used to expresses [sic] hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group“. Gongane et al. (2022, S. 3) bezeichnen Hatespeech als „detrimental content“, also schädliche Inhalte, die mit der Intention, eine bestimmte Person oder eine ganze Community anzugreifen, veröffentlicht werden. Angegriffen werden die Opfer

4 Siehe https://transparency.fb.com/data/community-standards-enforcement/ [Online, Zugriff am 24.11.2022].
5 Siehe https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec [Online, Zugriff am 24.11.22].
6 Siehe https://transparencyreport.google.com/youtube-policy/removals [Online, Zugriff am 24.11.22].

 DOI: 10.15460/apimagazin.2023.4.2.147

auf Basis ihrer Identität, zum Beispiel wegen ihres Geschlechts, Sexualität, Herkunft oder ihrer Hautfarbe (vgl. Davidson et al. 2017, S. 512). Siapera und Viejo-Otero (2021, S. 113) stellen gesondert heraus, dass Hatespeech, sowohl analog als auch im digitalen Raum, als Instrument zu verstehen ist, mit dem Rassismus ausgeübt wird. Hatespeech ist im Internet weit verbreitet und alltäglich präsent. Sowohl durchschnittliche Nutzer*innen, als auch extremistische Gruppen können Urheber*innen sein. Schlussendlich ist das Internet lediglich ein neues Medium für Hassrede, die es in gleichen oder ähnlichen Formen schon lange gibt (vgl. Siapera & Viejo-Otero 2021, S. 115f.). Zum Unterbinden und Entfernen von Hatespeech werden automatisierte Filter eingesetzt. Facebook beispielsweise nutzt Filter, die textbasierte Inhalte klassifizieren und einen Score vergeben, der aussagen soll, ob es sich um Hassrede handelt oder nicht. Diese automatisch geflaggten Inhalte werden von Moderator*innen final begutachtet, während YouTube auf vollautomatisierte Verfahren setzt (vgl. Gorwa et al. 2020, S. 9f.). Im Folgenden werden Probleme vorgestellt, die im Zusammenhang von Hatespeech und KI-Einsatz auftreten können. Im Nachgang werden Probleme genannt, die während dieses Prozesses entstehen.

4.1 Komplexität von Sprache

Das Erkennen von Hatespeech auf Social Media durch automatisierte Filter ist kein einfaches Unterfangen. Ein zentraler Grund hierfür ist die Komplexität von Sprache. Algorithmen können nicht so wie Menschen den Kontext einer Aussage erfassen. Durch Blacklists von Schlagwörtern kann die KI Inhalte missverstehen, die beispielsweise sarkastisch gemeint sind (vgl. Dachwitz & Reuter 2019, S. 60). Im Kontext von Hatespeech ist die Klassifizierung teilweise eine Gratwanderung zwischen Beleidigung und Hassrede. Davidson et al. (2017, S. 515) fanden in ihrer Untersuchung von 25.000 Tweets heraus, dass automatisierte Systeme, die auf Basis von einer Liste mit Schlüsselwörtern arbeiten, Schwierigkeiten hatten, subtilere Fälle von Hatespeech zu erkennen, wenn kein typischer Slur, also ein beleidigender und diskriminierender Ausdruck, enthalten war. Außerdem wurden Tweets, die zwar einen Slur enthielten, aber durch ihren Kontext nicht als Hatespeech zu verstehen waren, fälschlicherweise als solche eingeschätzt.

Selbst Systeme, die Natural Language Processing nutzen und damit als fortschrittlicher gelten, können nicht das gesamte Spektrum von Sprache begreifen (vgl. Dias Oliva et al. 2021, S. 704). Auch kulturelle sprachliche Besonderheiten oder die Verwendung von Schimpfwörtern unter Freund*innen kann eine KI schwer interpretieren (vgl. Cobbe 2021, S. 740). Algorithmische Filter zur Erkennung von schädlicher Sprache, wie zum Beispiel Perspective, konnten in Studien nicht immer Zuverlässigkeit beweisen (vgl. Gorwa et al. 2020, S. 10). Selbst wenn semi-automatisierte Verfahren angewendet werden, also der von der KI gekennzeichnete Inhalt noch einmal von einem Menschen geprüft wird, kann es zu Fehlentscheidungen kommen, wenn dem oder der Moderator*in ebenfalls der nötige Kontext zur Beurteilung fehlt (vgl. ebd., S. 10). So werden Inhalte unrechtmäßig entfernt, die zum Beispiel als Satire

DOI: 10.15460/apimagazin.2023.4.2.147

von der Kunstfreiheit oder als Meinungsäußerung geschützt sind (vgl. Löber 2022, S. 292).

4.2 Unrechtmäßiges Entfernen von Inhalten und Diskriminierung

Der vorangegangene Punkt ist eng mit diesem verknüpft. Die fehlerhafte Klassifizierung von Inhalten als Hatespeech kann besonders diejenigen treffen, die die Ziele des Hasses sind. Besonders bemerkbar macht sich dieses Problem zum Beispiel in der queeren Community. Kulturell ist es für queere Menschen üblich, sich Slurs anzueignen. Es findet sozusagen eine Zurückeroberung der unterdrückenden Worte statt. Es werden Schimpfwörter benutzt, die untereinander als liebevoll interpretiert werden, während sie von Nichtzugehörigen der Gruppe als Angriff zu werten sind (vgl. Dias Oliva et al. 2021, S. 706). Online hat das zur Folge, dass diese Inhalte sanktioniert werden. Dias Oliva et al. (2021, S. 712f.) stellten in einer Untersuchung fest, dass Twitter-Accounts von Drag Queens von Perspective, der bereits erwähnten KI von Jigsaw aus dem Google-Umfeld, im Durchschnitt als schädlicher eingestuft wurden als die Accounts von Donald Trump und anderen weißen Nationalist*innen. Das lag vor allem daran, dass Drag Queens häufiger Schimpfwörter wie „bitch“ nutzen, dies aber liebevoll konnotiert ist. Außerdem stufte die KI Worte wie „gay“ oder „lesbian“ per se als schädlich ein. Denn häufig von weißen Nationalist*innen genutzten Worte wie „white“, „black“, oder „Christian“ ordnete die KI ein geringeres Maß an Schädlichkeit („toxicity“) zu, obwohl sie im Gesamtkontext als diskriminierend gewertet werden sollten (vgl. ebd., S. 722). Das Unvermögen künstlicher Intelligenz, Kontext und Kultur zu verstehen, kann also dazu führen, dass Inhalte von Angehörigen marginalisierter Gruppen zu Unrecht weniger sichtbar sind, weil sie fälschlicherweise als Hassrede eingeordnet werden, und so das Recht auf freie Rede beschnitten wird (vgl. ebd., S. 729).

Zu ähnlichen Ergebnissen gelangten auch Haimson et al. (2021). Die Forscher*innen betrachteten drei Gruppen, von denen oft Vorwürfe der überproportionalen Sperrung von Accounts und Content entspringen: (politisch) Konservative, Trans-Menschen und Angehörige ethnischer Minderheiten (vor allem Schwarze Menschen). Ihre Umfrage zeigte, dass alle drei Gruppen ähnlich stark von Löschungen betroffen waren, nur dass sich die Art der Inhalte unterschied. Während es sich bei den entfernten Inhalten von Konservativen meist um Falschinformationen oder beleidigende Sprache bis hin zu Hassrede handelte, also um Inhalte, die gegen Plattformregeln verstoßen, zeigte sich bei Schwarzen und trans Nutzer*innen ein anderes Bild. Ihre gelöschten Inhalte behandelten ihre queeren beziehungsweise schwarzen Identitäten: Oft ging es um Erfahrungen mit Diskriminierung und Gerechtigkeit. Dieser Content wurde entfernt, obwohl er augenscheinlich den Plattformregeln folgte oder zumindest in eine Grauzone fiel (vgl. Haimson et al. 2021, S. 22f.). Content Moderation beschränkte diese Gruppen also im Ausleben ihrer Identität und in der öffentlichen Diskursteilhabe (vgl. ebd., S. 27). Nicht klar festzustellen ist bei dieser Studie, wie viel Einfluss der Einsatz von KI auf die Moderationsentscheidungen hatte, da als

 DOI: 10.15460/apimagazin.2023.4.2.147

Methode die Befragung von Betroffenen gewählt wurde. Da automatisierte Filter aber auf allen großen sozialen Plattformen gängig sind, ist davon auszugehen, dass diese auch eine Rolle spielten.

4.3 Besonderheit von Rassismus

Die großen sozialen Plattformen Facebook, Twitter und YouTube haben ihre Umgangsweisen mit Hatespeech grob nach den herrschenden rechtlichen Vorgaben ausgerichtet (vgl. Siapera 2022, S. 57). Siapera betrachtet diese Richtlinien und ihren Umgang mit Rassismus im Speziellen aus einer dekolonialen Perspektive. Sie zeigt auf, dass die Strukturen von KI-gestützter Content Moderation eine Spiegelung von bereits bestehenden Machtstrukturen sind, wie sie zum Beispiel im Justizsystem zu finden sind. Rassismus wird also nicht als strukturelle Macht erkannt. Das ist Siaperas zentraler Kritikpunkt an den Beschreibungen von Hatespeech in den Guidelines verschiedener Plattformen: die Eigenheiten und die spezielle Geschichte rassistischer Diskriminierung gegenüber anderer Diskriminierungsformen wird ignoriert. Kritisches Weißsein und eine Distinktion zwischen Unterdrückenden und Unterdrückten findet nicht statt (vgl. ebd., S. 58). Darüber hinaus unterstreicht sie auch die mangelhafte Einbindung von rassistisch diskriminierten Menschen in die Gestaltung von Content Moderation sowie die bereits erläuterten ausbeutenden Arbeitsbedingungen für Moderator*innen (vgl. ebd., S. 57). Siapera fordert, dass die Erfahrungen von den Menschen, die von rassistischer Sprache betroffen sind, erfragt werden sollten. Außerdem sollten Moderator*innen spezialisierte Trainings zu rassistischem Hass bekommen. Denn ihre Moderationsentscheidungen bilden die Trainingsdaten für den KI-Einsatz ab (vgl. ebd., S. 61). Die Gleichsetzung von verschiedenen Diskriminierungsformen im Feld der Content Moderation wird auch von Marshall (2021, S. 6) kritisiert. Die Gefahr, dass automatisierte Systeme durch diese fehlende Sensitivität bestehende ungerechte gesellschaftliche Normen fortführen, ist gegeben (vgl. Gorwa et al. 2020, S. 11). Siapera und Viejo-Otero (2021, S. 127) stellen bei der Betrachtung von Facebooks automatisierter Content Moderation fest, dass keine Hinterfragung von rassistischen Machtstrukturen stattfindet. Stattdessen wird Rassismus als eine von mehreren Arten von Hatespeech behandelt. Die Autor*innen führen an, dass Hatespeech von Facebook nicht aus ethischen oder politischen Gründen entfernt wird, sondern als Inhalt, der sich negativ auf die Interaktion und Sicherheit von Nutzer*innen und somit auf die Attraktivität der Plattform auswirkt. Es wird bemerkt, dass Facebook sogar von Hatespeech profitiert, denn das Entfernen von Hatespeech produziert Trainingsdaten für die KI, die die Content Moderation übernimmt. Mit der KI spart Facebook Geld, da menschliche Moderation kostenintensiver ist. All diese Dynamiken führen aber nicht zu einer nachhaltigen Beschäftigung mit und Minimierung von rassistischen Inhalten, so die Autor*innen.

4.4 Objektivität und Bias

Im Kontext von KI-gestützter Content Moderation stellten Molina und Sundar (2022, S. 14 f.) fest, dass Vielnutzer*innen und Nutzer*innen mit einer generellen Angst vor

DOI: 10.15460/apimagazin.2023.4.2.147

künstlicher Intelligenz skeptisch sind und glauben, dass eine KI keine nuancierte Moderationsentscheidung treffen kann. Gleichzeitig glaubten Nutzer*innen, die kein Vertrauen in andere Menschen haben sowie politisch konservativ eingestellte Nutzer*innen, dass eine KI eine objektivere und richtigere Moderationsentscheidung als ein Mensch fällen kann. Trotzdem kann ein Ausbleiben von falschen Entscheidungen einer KI nicht garantiert werden (vgl. Löber 2022, S. 292). Ein Bias kann ebenso in Algorithmen vorkommen, so auch im Feld der Content Moderation. Der Bias kann von Entwickler*innen stammen oder auch von den Trainingsdaten. Diese werden direkt durch Entscheidungen von menschlichen Moderator*innen beeinflusst, die natürlich auch zu einem gewissen Grad subjektiv sind. Annotierte Trainingsdaten (siehe Punkt 2.2.) sind zwar gut geeignet, um Hatespeech zu erkennen, die Kennzeichnungen können aber ebenso einen Bias und Fehler enthalten (vgl. Llansó et al. 2020, S. 4). Auch wenn Inhalte nicht von einer KI, sondern von Nutzer*innen geflaggt werden, kann ein „Popularity Bias" (vgl. Haimson et al. 2021, S. 3) auftreten, wenn die Mehrheit der Nutzer*innen Normen vertreten, die marginalisierten Gruppen schaden. Es ist unklar, ob der Einsatz von KI gerechtere Content Moderation nach sich zieht (vgl. Wang & Zhu 2022, S. 825).

Dieser Punkt steht in Verbindung zum Punkt Transparenz (unter 3.1.). Mangelnde Transparenz bietet Raum für Vermutungen und Verschwörungserzählungen. So neigen politisch Konservative zu der Behauptung, dass ihre Inhalte auf sozialen Plattformen eher entfernt werden als andere, dass es also einen politischen Bias gäbe (Jiang et al. 2020, S. 13669). Jiang et al. (2021, S. 13771) stellten in ihrer Studie fest, dass Kommentare unter Videos mit politisch eher rechten Inhalten zwar tatsächlich häufiger entfernt werden, dass es dafür aber auch berechtigte Gründe gibt. Die Kommentare verstießen häufiger gegen Plattformregeln, indem sie beispielsweise Hatespeech oder Beleidigungen enthielten. Ein politscher Bias konnte nicht nachgewiesen werden.

4.5 Gesetze und Interessen einzelner Nationen

So wie es keine allgemeingültige Definition von Hatespeech gibt, ist auch die Gesetzeslage zu deren Regulierung global nicht einheitlich. Hier zeigt sich, dass ein Regelwerk einer Plattform keine global einheitliche Formulierung sein muss. Facebook zum Beispiel verfährt bei Holocaustleugnung nicht einheitlich: während solche Inhalte in Ländern wie Deutschland, wo die Holocaustleugnung illegal ist, geblockt werden, kann man sie in anderen Teilen der Welt weiterhin sehen. Diese Praxis wird auch Geoblocking genannt (vgl. Petricca 2020, S. 315). Nach einem Urteil des Europäischen Gerichtshofs aus dem Jahre 2019[7] ist Facebook allerdings gezwungen, Inhalte global zu entfernen, wenn irgendein Gericht eines EU-Landes diesen Inhalt als illegal eingestuft hat. Hier reicht Geoblocking dann nicht mehr aus. Kritiker*innen bemängeln, dass so die Meinungsfreiheit gefährdet sein könnte, wenn die Gesetze

7 Siehe: https://curia.europa.eu/juris/document/document.jsf?text=&docid=218621&pageIndex=0&doclang=DE&mode=req&dir=&occ=first&part=1&cid=1956673 [Online, Zugriff am 16.06.2023].

DOI: 10.15460/apimagazin.2023.4.2.147

einer einzelnen Nation so global wirken (vgl. ebd., S. 323). Generell herrscht international Uneinigkeit im rechtlichen Umgang mit digitaler Hassrede. Zusätzliche Schwierigkeiten entstehen, weil das Internet die Möglichkeit bietet, solche schädlichen Inhalte über Ländergrenzen hinaus zu verbreiten und dabei zudem anonym zu bleiben (vgl. Lee & Li 2020, S. 2). Fälle wie diese zeigen, dass Content Moderation und die Gestaltung von Algorithmen dynamisch sind.

Das bereits erwähnte, in Deutschland gültige NetzDG (Netzwerkdurchsetzungsgesetz) beeinflusst und forciert den Einsatz von künstlicher Intelligenz in der Content Moderation. Es wurde vom DSA (Digital Services Act) abgelöst, der am 16.11.2022 in Kraft trat. Sozialen Plattformen drohen sehr hohe Strafen, wenn Inhalte, die gegen das Gesetz verstoßen, nicht innerhalb von 24 Stunden entfernt werden (vgl. Petricca, S. 322). Dieser zeitliche Rahmen lässt sich angesichts der Masse an Content nur mit automatisierten Systemen einhalten. Das NetzDG bewirkte, dass seit seinem Inkrafttreten eine größere Anzahl an Inhalten entfernt und kontrolliert wird, vor allem im Bereich Hatespeech und Beleidigung. Allerdings lässt sich daraus nicht schließen, ob Content Moderation durch diesen staatlichen Eingriff fairer geworden ist (vgl. ebd., S. 323).

Doch nicht nur offizielle, feststehende Gesetze haben einen Einfluss. Auch politischen Interessen in Krisen geht Facebook teilweise nach. Das macht Content Moderation und ihre Algorithmen umso undurchsichtiger. Die Entscheidungen automatisierter Filter werden augenscheinlich nicht nur von dem Regelwerk beeinflusst, welches sie umsetzen, sondern auch von den politischen Interessen des Landes, in dem sie operieren. Petricca (2020, S. 316) liefert dafür ein Beispiel: Nutzer*innen von Google Maps, die die Website aus Russland öffneten, wurde die Krim als Teil russischen Gebiets angezeigt, während Nutzer*innen aus der restlichen Welt die Halbinsel als Teil der Ukraine dargestellt wurde. Hier wird deutlich, wie Plattformen in aktuellen Krisen navigieren müssen.

5 Fazit

In dieser Arbeit wurde der Themenkomplex automatisierte Content Moderation und Hatespeech beleuchtet. Bei der Betrachtung der Probleme, die mit KI-gestützter Content Moderation im Allgemeinen und beim Umgang mit Hatespeech auf sozialen Plattformen im Speziellen auftreten können, wurde deutlich, dass die Aufgabenstellung Content Moderation ein sehr komplexes Unterfangen für die Plattformbetreiber*innen ist. Probleme wie mangelnde Transparenz, die Einstufung von KI-gestützter Content Moderation als automatisierte Zensur und die daraus folgende Frage, welche Rechenschaftspflicht es für Plattformbetreiber*innen gibt, sind zentrale Bedenken in der wissenschaftlichen Literatur.

DOI: 10.15460/apimagazin.2023.4.2.147

Die erläuterten Punkte im Themenkomplex automatisierte Content Moderation und Hatespeech zeigen, dass eine KI nicht dem Prinzip „one size fits all“ folgen kann. Kulturelle Gegebenheiten, geopolitische Interessen, Gesetze und Bedürfnisse verschiedener marginalisierter Gruppen stellen Einflüsse auf die Gestaltung dieses Themenkomplexes dar. Allein die Tatsache, dass es keine allgemeingültige Definition von Hatespeech gibt, spiegelt das wider. Jede Plattform, jede Nation und jede Subkultur kann verschiedene Standards im Umgang mit Hatespeech setzen. Während manche das automatisierte Entfernen von schädlichen Inhalten als nötige Regulierung von Hass sehen, bezeichnen andere es als eine Beschneidung der Meinungsfreiheit (vgl. Siapera & Viejo-Otero 2021, S. 115). Die Entwicklung von Strategien über das bloße Entfernen von Hatespeech hinaus ist wünschenswert. Wie kann Hatespeech und Diskriminierung nachhaltig minimiert werden, zum Beispiel durch Content Moderation mit anschließender Vermittlung von Bildungsmaterialien (vgl. Lee & Li 2020, S. 4)? Die achtsame Gestaltung von Content Moderation sowie Formulierung von Regeln ist von großer Bedeutung: Denn „platform moderation is never articulated in a vacuum“ (Zolides 2021, S. 3000). Regelwerke, die von automatisierten Systemen umgesetzt werden, können herrschende Machtdynamiken in Bereichen wie Rassismus, Geschlecht und allen anderen Diskriminierungsformen spiegeln und verstärken (vgl. ebd., S. 3013). Das hat schlussendlich Auswirkungen auf das Leben und die Kommunikation von marginalisierten Menschen (vgl. Marshall 2021, S. 12). Forschung zur Vermeidung von Bias in der Content Moderation und den dafür verwendeten Algorithmen ist deshalb sinnvoll.

Der Ausblick auf die Zukunft zeigt kein klares Bild. Facebook könnte in den nächsten zehn Jahren komplett auf automatische Filter umstellen und somit menschliche Moderator*innen überflüssig machen (vgl. Dachwitz & Reuter 2019, S. 61). Fraglich ist, ob das angesichts der Probleme, die in dieser Arbeit illustriert wurden, wünschenswert ist, beziehungsweise ob sich sinnvolle Lösungen für diese Probleme finden werden. Gillespie (2020, S. 4) hinterfragt in dem Zusammenhang, ob es überhaupt eine gut funktionierende Form der Content Moderation geben kann. Der Einsatz von KI, der die beschriebenen Probleme mit sich bringt, wird mit der Größe der Plattformen und der Masse an Inhalten gerechtfertigt, die menschlich nicht mehr bearbeitbar ist. Vielleicht gibt es einen Punkt, an dem Plattformen schlicht zu groß und nicht mehr sinnvoll kontrollierbar werden, auch nicht mit automatisierten Systemen. Hier stößt die kapitalistische „‘growth at all costs‘ mentality“ unter Umständen an ihre Grenzen. Ganz ohne künstliche Intelligenz wird allerdings auch niemand auskommen können. Neben der Masse an Content ist vor allem die psychische Gesundheit der Moderator*innen ein wichtiges Argument für den Einsatz von automatisierten Filtern. Diese können die Moderator*innen davor bewahren, tagtäglich Unmengen an Gewaltdarstellungen und extremistischer Propaganda ausgesetzt zu sein, und stellen somit eine ethisch vertretbare Weise dar, Content Moderation zu bewältigen (vgl. ebd., S. 4).

DOI: 10.15460/apimagazin.2023.4.2.147

Zusammenfassend lässt sich festhalten, dass die Regulierung von Hatespeech auf Social Media durch automatisierte Filter einer Reihe von Schwierigkeiten gegenübersteht. Eine zentrale Rolle spielt dabei die Komplexität und Dynamik der menschlichen Sprache, die es verlangt, künstliche Intelligenzen und Regelwerke fortlaufend anzupassen und weiterzuentwickeln. Für diesen Prozess ist es wünschenswert, dass Angehörige marginalisierte Gruppen befragt und eingebunden werden, um den Besonderheiten verschiedener Diskriminierungsformen Sorge zu tragen.

DOI: 10.15460/apimagazin.2023.4.2.147

Literatur

BARRETT, P. M., 2020. Who moderates the social media giants? A call to end outsourcing. NYU Stern Center Centre for Business and Human Rights. https://bhr.stern.nyu.edu/tech-content-moderation-june- 2020?_ga=2.253758285.418266349.1669401682-1382939569.1669131382 [abgerufen am 03.11.2022]

COBBE, J., 2021. Algorithmic censorship by social platforms: Power and resistance. Philosophy & Technology, 34(4), 739-766 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1007/s13347-020-00429-0.

DACHWITZ, I., und REUTER, M., 2019. Warum Künstliche Intelligenz Facebooks Moderationsprobleme nicht lösen kann, ohne neue zu schaffen. FIfF-Kommunikation, 19(2), 59-61.

DAVIDSON, T., WARMSLEY, D., MARCY, M., und WEBER, I., 2017. Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.48550/arXiv.1703.04009.

DIAS OLIVIA, T., ANTONIALLI, D. M., und GOMES, A., 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. Sexuality & Culture, 25(2), 700-732 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1007/s12119-020-09790-w.

DJEFFAL, C., 2022. Soziale Medien und Kuratierung von Inhalten. Regulative Antworten auf eine demokratische Schlüsselfrage. In I. Spiecker gen. Döhmann, M. Westland & R. Campos (Hrsg.), Demokratie und Öffentlichkeit im 21. Jahrhundert–zur Macht des Digitalen (S. 177-198). Nomos Verlagsgesellschaft.

EFRON, S. E., und RAVID, R., 2019. Writing the Literature Review: A Practical Guide. New York: Guilford Press.

GILLESPIE, T., 2020. Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 1-5 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/20539517209432.

GONGANE, V. U., MUNOT, M. V., und ANUSE, A. D., 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. Social Network Analysis and Mining, 12(1), 1-41 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1007/s13278-022-00951-3.

GORWA, R., BINNS, R., und KATZENBACH, C., 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 1-15 [Zugriff am 27.06.2023]. Verfügbar unter: 10.1177/2053951719897945.

GRIMMELMANN, J., 2015. The virtues of moderation. Yale JL & Tech., 17, 42.

HAIMSON, O. L., DELMONACE, D., NIE, P., und WEGNER, A., 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-35 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1145/3479610.

JIANG, S., ROBERTSON, R. E., und WILSON, C., 2020. Reasoning about political bias in content moderation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(9), 13669-13672 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1609/aaai.v34i09.7117.

JIMENEZ DURAN, R., 2022. The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. Social Science Research Council; University of Chicago. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4044098 [abgerufen am 02.11.2022]

KOR-SINS, R., 2021. The alt-right digital migration: A heterogeneous engineering approach to social media platform branding. New Media & Society, 1-18 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/14614448211038810.

LEE, R. K. W., und LI, Z.,2020. Online xenophobic behavior amid the COVID-19 pandemic: a commentary. Digital Government: Research and Practice, 2(1), 1-5 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1145/3428091.

LLANSO, E. J.,2020. No amount of „AI“ in content moderation will solve filtering's prior- restraint problem. Big Data & Society, 7(1), 1-6 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/2053951720920686

LLANSO, E., VAN HOBOKEN, J., Leerssen, P., und HARAMBAM, J., 2020. Artificial intelligence, content moderation, and freedom of expression. Transatlantic Working Group on Content Moderation Online and Freedom of Expression. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf [abgerufen am 02.11.2022]

LÖBER, L. I.,2022. KI-Lösungen gegen digitale Desinformation: Rechtspflichten und-befugnisse der Anbieter von Social Networks. In M. Friedewald, A. Roßnagel, J. Heesen, N. Krämer, J. LAMLA, (Hrsg.), Künstliche Intelligenz, Demokratie und Privatheit (S. 289- 316). Nomos Verlagsgesellschaft.

MA, R., und KOU, Y., 2021. „How advertiser-friendly is my video?“: YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-25 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1145/3479573.

MARSHALL, B., 2021. Algorithmic misogynoir in content moderation practice. Heinrich- Böll-Stiftung European Union. https://eu.boell.org/sites/default/files/2021-06/HBS-e-paper-Algorithmic-Misogynoir-in-Content-Moderation-Practice-200621_FINAL.pdf [abgerufen am 05.11.2022]

MOLINA, M. D., und SUNDAR, S. S., 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. Journal of Computer-Mediated Communication, 27(4), 1-12 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/14614448221103534.

OTTER, D. W., MEDINA, J. R., und KALITA, J. K., 2020. A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), 604-624 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.48550/arXiv.1807.10854.

PETRICCA, P., 2020. Commercial Content Moderation: An opaque maze for freedom of expression and customers' opinions. Rivista internazionale di Filosofia e Psicologia, 11(3), 307-326 [Zugriff am 27.06.2023]. Verfügbar unter: 10.4453/rifp.2020.0021.

PINCHEVSKI, A., 2022. Social media's canaries: Content moderators between digital labor and mediated trauma. Media, Culture and Society, 45(1), 212-221 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/01634437221122226.

SIAPERA, E., 2022. AI Content Moderation, Racism and (de) Coloniality. International Journal of Bullying Prevention, 4(1), 55-65 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1007/s42380-021-00105-7.

SIAPERA, E., und VIEJO-OTERO, P., 2021. Governing hate: Facebook and digital racism. Television & New Media, 22(2), 112-130 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/1527476420982232.

DOI: 10.15460/apimagazin.2023.4.2.147

WANG, L., und ZHU, H., 2022. How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. 2022 ACM Conference on Fairness, Accountability, and Transparency (S. 824-838). https://dl.acm.org/doi/abs/10.1145/3531146.3533147 [abgerufen am 18.11.2022]

WRIGHT, L., 2022. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. Social Media+ Society, 8(1), 1-11 [Zugriff am 27.06.2023]. Verfügbar unter: DOI: 10.1177/20563051221077020.

ZOLIDES, A., 2021. Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. New Media & Society, 23(10), 2999-3015 [Zugriff am 27.06.2023]. Verfügbar unter: 10.1177/1461444820942483.